MODELING INTERACTIONS OF FLEXIBLE PROTEINS

by

Shourya S. Roy Burman

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

October, 2018

© Shourya S. Roy Burman 2018

All rights reserved

Abstract

Proteins are dynamic molecules that mediate most biological processes through interactions with other proteins and biomolecules. A fundamental understanding of the mechanisms governing protein interactions requires intricate knowledge of the three-dimensional structures of biomolecular complexes. Despite advances in experimental structure determination, we have structural insights into only a small fraction of known complexes. Computational modeling provides an invaluable complementary tool to explore protein interactions in a rapid and high-throughput manner. A principal challenge limiting the accuracy of current computational methods is the ability to predict binding-induced conformational changes during protein–protein association. In this dissertation, I address this challenge by creating new tools to predict atomistic models of flexible protein complexes. First, I develop a heterodimer docking protocol that incorporates flexibility by efficiently simulating conformational selection from hundreds of pre-generated backbone conformations and identifies the near-native models with a novel, coarse-grained score function called Motif Dock Score (MDS). On a benchmark of 88 complexes with different degrees of flexibility, this protocol, RosettaDock 4.0, is the first method to successfully dock approximately 50% of complexes with conformational change of up to 2.2 Å. Next, I present the results of our participation in the community-wide blind experiment, Critical Assessment of PRedicted

ABSTRACT

Interactions (CAPRI) rounds 37–45, where I use various docking methods to predict the structures of protein homomer, heteromer and oligosaccharide complexes. In the process, I identify inadequacies in these methods and propose enhancements. Based on the shortcomings identified in CAPRI, I develop a protocol to predict the structure of symmetric homomers from monomeric inputs with a focus on tightly-packed complexes. This method, Rosetta SymDock2, leverages MDS in the coarse-grained phase and simulates subunit flexibility through induced fit by all-atom flexible-backbone refinement. It outperforms competing algorithms by docking 61% of cyclic complexes and 42% of dihedral complexes in a diverse benchmark of 43 homomers. In the course of developing these algorithms, I also discover that the binding energy wells of homomers are narrower, steeper and deeper than those of heterodimers, thus explaining their increased stability. Finally, I present preliminary results to propose data-driven strategies that can overcome current barriers to accurate modeling.

Advisor: Prof. Jeffrey J. Gray (Department of Chemical & Biomolecular Engineering)

Reader: Prof. Rebecca Schulman (Department of Chemical & Biomolecular Engineering)

Reader: Prof. Margaret E. Johnson (Department of Biophyphysics)

Acknowledgements

This work would not have been possible without the help and support of a large number of people. First-and-foremost, I would like to thank my advisor, Professor Jeffrey Gray for guiding me through six years of graduate school. I had always wanted to work in computational biology, but was slightly apprehensive about making the leap to computational method development in Jeff's lab. During advisor selection, while walking back from class Professor Joelle Frechette gave me the final push by saying, "If you're remotely interested in that sort of thing, I can tell you that people in Jeff's lab are really happy." As expected, I found the learning slope of Rosetta to be quite steep and barely managed to read the complex codebase, let alone write new code. To ease me into the lab, Jeff leveraged my prior benchwork experience on a high-risk, high-reward project on biomineralization. Despite it not being a core focus of his lab, he ensured that I got trained in state-of-the-art imaging and characterization techniques by experts in the respective fields. In due course, I had gained enough experience to seamlessly transition into developing new Rosetta protocols and he encouraged me to do so.

Right from the very beginning, Jeff asked me to read whatever I found interesting and to form project ideas based on it. It took me many years and many missteps to appreciate the importance of this creative freedom to develop critical thinking. He also requested me to write parts of grant proposals with him, which gave me a firsthand view of the process. Further, he

helped me improve as a writer by constantly critiquing my style and syntax. Besides professional development, Jeff also focused on diversity and inclusion, areas about which I did not think as much before. By associating with him, I learnt about people management and how to maximize the output from a group with a diverse set of skills and life experiences. The collaborative and respectful nature of our group, with members past and present, echoes Jeff's attitude. To us, he is as much of a friend as he is a mentor.

I would like to thank Professors Rebecca Schulman, Margaret Johnson, Yannis Kevrekidis, Jungsan Sohn, and Marc Ostermeier for serving on my thesis committee. Over the years, Professors Schulman and Johnson have given me invaluable feedback, such as how I should focus on discriminating near-native models and not just generating them.

My work was funded by the National Institutes of Health (grant no. R01-GM078221) and the National Science Foundation (award no. 1507736). My tuition and stipend have been borne in part by the Johns Hopkins University. The Maryland Advanced Research Computing Center and the Texas Advanced Computing Center have generously provided me with millions of CPU hours to run my simulations.

Throughout my stay, the Gray Lab always had an excellent work environment. Michael Pacella was my first mentor and collaborator in the lab. His designs on biomineralization formed the basis of my experiments. As I dug deeper into the Rosetta codebase, my other office mate, Jason Labonte became my *de facto* mentor. When I started developing protocols, I would pester him with numerous questions every day and ask for his opinion on a plethora of bug fixes. Although we had a brief overlap, Daisuke Kuroda's work on characterizing

ensemble generation methods was my inspiration and go-to resource to improve sampling for flexible backbone proteins. While I was focusing on improving sampling, based on prior work by William Sheffler at the University of Washington, Nicholas Marze optimized this wondrous scoring scheme. By combining his score function with my sampling strategies, Nick and I created RosettaDock 4.0. After his graduation, I continued to use his scoring technique to develop Rosetta SymDock2. Remy Yovanno, a rotation student was instrumental in designing this protocol and testing it initially.

A unique experience in the Gray lab is participating in a blind prediction challenge called CAPRI. When I first joined the prediction crew, I learnt the ropes under Nick's and Daisuke's guidance. Over the years, I had the opportunity to work with several team mates who have taught me new techniques and tricks, none more so than Jeliazko Jeliazkov. Besides adopting his style of maintaining records, I have followed his lead on loop modeling and antibody docking. In the last few months, Morgan Nance ably stepped into my role as the CAPRI lead, allowing me more time to develop new methods. Jason and Morgan also helped me understand and model oligosaccharide–protein interactions. Besides them, I have been fortunate to work with Joseph Lubin and Naireeta Biswas on some targets. Modeling is greatly improved when we have experimental data to complement our predictions. For one target, I consulted Prof. Jamie Spangler, who not only discussed the problem, but also shared a rough model that she had come up with during her studies, which helped me correctly predict that target. I would be remiss not to thank the CAPRI organizers, Professors Shoshana Wodak, Sameer Velankar, and Marc Lensink for painstakingly putting together the targets and evaluating our predictions.

ACKNOWLEDGEMENTS

I have often sought advice on both research and career matters from Krishna Kilambi, most recently, when I was trying to choose between an academic and industrial postdoc. I am grateful to Julia Koehler Leman for including me in the recent grand review of Rosetta methods. I have had wonderful conversations with Xiyao Long, who introduced me to authentic Chinese cuisine and garb. Kavyon Tabrizi, a surprisingly gifted undergraduate student in our lab, constantly asked some very fundamental questions and made me revisit a lot of mathematics and physics I had long forgotten. I have spent countless hours discussing science and everything else with several other Gray lab members including Brian Weitzner, Rebecca Alford, Elizabeth Lagesse, Sai Pooja Mahajan, and Jing Zhou. I have also been extremely lucky to have several capable mentees like Remy, Drew Morley, Kathy Wang, Nikhil Shah, and Kathleen Rand. Two people crucial to the design of my work were Sergey Lyskov and Matt Mulqueen, both of whom I bombarded with technical questions and who greatly helped me improve my practical knowledge of computer science.

I have been fortunate to get a number of opportunities to apply my computational skills to real world applications through several collaborations that are not discussed in this dissertation. In my first two years, Professor James De Yoreo and Jinhui Tao helped me learn atomic force microscopy on their instrument at the Pacific Northwest National Lab. Realizing that I was living in isolated quarters without a means of transportation, Jinhui and his wife would drive me to the grocer's on the weekends. Professor Philip Cole and Jay Kalin introduced me to rational drug design, and with Jeli's help, I proposed some plausible models for their system. A stimulating problem to test my flexible docking protocols was proposed

by Professors Cynthia Wolberger and Patrick Lombardi, with whom I am still collaborating. Professor Wolberger also advised me on selecting my postdoctoral advisor. Another ongoing project on enzyme design, which has been tremendously fun to work on, is in collaboration with Professor Steven Rokita and Zuodong Sun. Professors Rokita and Schulman also wrote letters of reference for me during my job search, for which I am eternally grateful.

I have sought advice on designing protocols and benchmarks from Professors Ingemar André, Frank Di Maio, Andrew Leaver-Fay and David Baker. I often used the Robetta server for my work, and David Kim was always eager to help me design specialized runs and to upgrade my priority.

Throughout my stay at Johns Hopkins, a rotating cast of staff in our department has supported my work. Those who have helped me directly include Caroline Qualls, Carla Gourdine, Giselle Rejas, Porscha Reid, Tiara Carr, Lucy Raybon, Beth Rannie, and Kourtney Roussey.

Jason set up a long-running Dungeons & Dragons campaign, which many former members of our lab regularly participate in. I am glad that I got to be a part of it and in the process got acquainted with the weird and wonderful virtual realm of magic and monsters in the medieval times. Jason came up with the most fantastical plots and my fellow lab mates-cum-"explorers" Joey, Mike, Nick, Brian, Krishna, and Elizabeth filled in the details with hilarious twists and turns. While they were still around, D&D was pretty much all that was discussed during breaks, with some science and philosophy on the side. We were soon joined by new explorers David Ryoo and Jared Labonte to complete the team.

ACKNOWLEDGEMENTS

I have relied on a large network of friends in Baltimore and abroad to get through my years away from family. In particular, I had a great time with my roommates Varun, Akshay, Navaneethan, Sravya, and Kousik. We always thought on a similar wavelength about politics, religion, and most importantly, food. In the first couple of years, I had a ridiculous amount of fun participating in events with Abha, Vaibhav, Ashish, Debmalya, Piyush, and Aagam. Towards the later years, a defining experience was game nights with Siddharth, Vishwa, Kesha, Gowtham, Vivek, Nagaraj, Anand, Princy, Pooja, Akanksha, Purnima, Mithra, Harini, and Ramanujan. This was supplemented by the badminton and cricket groups with whom I spent several hours a week. Whenever in need or otherwise, I phoned my hostel wing mates from college and chatted for hours.

Finally, and most importantly, I would like to thank my family. My parents have always encouraged me to pursue whatever intrigued me without worrying about the financial consequences. They provided me with all that I needed to study in the best schools and colleges without ever dictating my choices. Every day while talking to them on the phone they make sure that I am happy doing what I do and have a healthy dose of recreation on the side. Their continual background presence over the years has ensured that I remain composed through some of the more trying periods during my Ph.D. They have been particularly liberal about my social life and have allowed me to treat them as friends and confidants as I stepped into adulthood. Most recently, they have welcomed Rujuta into our family with open arms. Although it has been only five years since Rujuta and I have been together, I find it difficult to remember a time without her. Her bubbly and spontaneous nature completely contrasted

my more measured and unemotional approach to things, and yet somehow, we have managed to navigate through life together. In the process, she has enriched me by showing me to how to associate with people who are not like-minded, pointing out the rigidity of my ideas, and of course, teaching me a lot of Bollywood trivia. Her initiative of learning my mother tongue, Bengali, has also inspired me to learn hers, Marathi, about which I have been lax lately. While I slowly finished my graduate studies, she displayed tremendous patience to wait for the next stage of our lives, and I truly appreciate that. Last, but not least, I would like to thank my cousins, aunts and uncles for those comforting phone and video calls.

To Ma, Baba & Ruju

Table of Contents

CONTENTS

CONTENTS

List of Tables

List of Figures

CONTENTS

Chapter 1

Introduction

1.1. Protein–protein interactions

Biological processes are governed by intricate interaction networks of proteins and other biomolecules. Although advances in genome sequencing have supplied comprehensive lists of gene products, large-scale annotation of the functional role of these biomolecules in interaction networks is in its infancy.[1] A fundamental understanding of the specific interactions requires detailed three-dimensional structures of biomolecular complexes. Experimental structure determination using techniques like x-ray crystallography, Nuclear magnetic resonance (NMR) spectroscopy, and cryogenic electron microscopy (cryoEM) is labor-intensive, low-throughput and simply impossible for certain complexes. Computational prediction of protein contacts offers a promising alternative for many experimentally intractable complexes.[2] The most structurally detailed of these methods 'dock' three-dimensional models of the interacting partners using physical laws and empirical observations.

The sheer scale on which protein structures can be computationally analyzed was recently shown by structurally annotating antibody repertoires comprising millions of sequences.[3] A pipeline that allows large-scale computational docking of antibody repertoires with target antigens can drastically reduce the number of validating experiments and accelerate the discovery of antibody therapeutics. The first step towards such a pipeline was demonstrated by cross-docking antibodies with cognate and non-cognate antigens and discriminating native interactions with moderate success.[4] Antibodies are just one class of proteins; large-scale docking has the potential to revolutionize the wider healthcare industry.

Another motivation to develop better docking methods is the design of biomolecular assemblies. Accurate prediction of protein–protein interactions in Rosetta—the computational framework central to my work—has enabled the design of proteins that self-assemble into custom macrostructures.[5–7] RosettaDock—the docking protocol enhanced in this work—has been used to design hetero-bifunctional ligands to dimerize proteins that do not naturally associate.[8] Most design applications require sub-angstrom accuracy in the prediction of the bound interface; in this dissertation, I present key advances towards this goal.

1.2. Computational modeling of protein interactions

Attempts to model protein interactions date back to the mid-1970s. Greer and Bush found that when they mapped the height and charge of the binding surfaces of the α and the β chains of methemoglobin onto grids and superimposed the grids, the respective pixels on the interacting surfaces had the exact opposite features in large blocks.[9] These blocks correspond

to the binding interface, and this study established that, at the interface, the interacting chains have shape and charge complementarity. Around the same time, Wodak and Janin performed the first automated search of how two proteins associate, giving rise to the field of protein–protein docking. They calculated orientations of trypsin inhibitor on trypsin that resulted in the highest number of intermolecular contacts.[10] Of the nine potential binding modes they identified, one was the biologically observed one. Janin and Wodak also demonstrated a practical use of computational docking: predicting the mechanism of hemoglobin quaternary structure change pathway upon binding oxygen. Using a similar search technique to the one employed for the trypsin–inhibitor complex, they showed that an α–β dimer of hemoglobin in the deoxygenated state can bind another dimer both the deoxygenated and the oxygenated states, but the oxygenated state preferentially bound another oxygenated dimer.[11] This bias forms the basis for cooperative oxygen binding in hemoglobin, which was eventually validated theoretically[12] and experimentally[13].

These early studies laid the foundation for today's computational docking efforts. The ideas of shape and charge complementarity at the interface are used by all state-of-the-art sampling and selection algorithms. Some creative startegies frequently borrowed from Janin's and Wodak's work are searching along a discretized translational and rotational space, using a simplified representation of protein side chains, and using symmetry to reduce the degrees of freedom. In 1986, Connolly formally defined the protein docking problem as: "Given the three-dimensional structures of any two proteins, is it possible to predict whether they will associate, and if so, in what way?"[14] In this dissertation, I presume that there this prior evidence

of biological association of the given proteins, and I explore schemes for predicting the three-dimensional structures of the complex.

1.2.1. Modern approaches to protein docking

Computational protein docking methods can be broadly divided into three classes: template-based modeling, free docking, and data-driven modeling. Template-based modeling methods mine the Protein Data Bank[15] to identify homologous complex templates, construct an initial model by copying backbones of the aligned protein fragments, and refine the model by building side chains and fixing loops and termini.[16] The underlying assumption is that protein folds are more conserved than sequences, and by extension, if a protein of fold f_1 binds a protein of fold f_2 in a certain orientation, all interacting f_1–f_2 protein pairs will have the same binding mode.[17] If structures of homologous complexes are not available or if the identified templates have a different binding mode, template-based modeling fails. With methods like cryoEM producing structures of previously inaccessible large assemblies and the emergence of curated homolog libraries,[18,19] the probability of finding a good template is ever-increasing.

Free docking methods search relative orientations of the given proteins to obtain a binding mode with the best shape and charge complementarity. The most prominent approach approximates a protein as a three-dimensional grid with a special representation for the surface and uses Fast Fourier Transform (FFT) search algorithms to exhaustively sample all orientations.[20–25] FFT-based docking offers an exceptionally fast enumeration of binding modes for rigid bodies, but struggles to account for protein flexibility as it operates through

immutable grids. Monte Carlo approaches that randomly perturb orientation and side-chain placement[26,27] and approaches that minimize the energy of the system along translational and rotational degrees[28] are also commonly used. While the latter methods are inefficient for global sampling, their forte is exploring local space based on some preliminary estimate of the binding region.

Data-driven modeling approaches rely on spatial restraints obtained from experimental procedures or bioinformatics predictions.[29–31] While data-driven approaches necessarily require this information, integrating experimental data in the form of restraints improves the accuracy of template-based and free docking methods as well. Some experiments commonly used to complement computational docking are NMR,[32] small-angle X-ray scattering,[33] Förster resonance energy transfer,[34] cross-linking,[35] and Hydrogen/Deuterium Exchange Mass Spectrometry.[36] Alternatively, *in silico* approaches can help refine experimental structures by fitting to electron density maps from cryoEM.[37,38] The key to modeling the most challenging complexes lies in ability to combine a variety of experimental data with the speed of computational search.

1.2.2. Contemporary challenges in protein–protein docking

Some of the significant challenges faced by all of the aforementioned methods are predicting water molecules that mediate interface interactions, estimating conformational changes due to binding, sampling all possible the relative orientations while accounting for flexibility, and modeling multi-component complexes.[39–45] Water-mediated interactions

(Figure 1.1A) are believed to influence the kinetics and the thermodynamics of association,[46] but a dearth of high-resolution PDBs with structured interface waters has meant that our understanding of their role in binding is still in its infancy.[47]

The most pressing concern in docking is the repeated failure to deal with binding-induced large structural changes despite years of research (Figure 1.1B–C).[48] If a homolog is available, template-based modeling can circumvent the need to predict conformational deviations by stitching individual segments together based on a bound template.[49] However, how physically realistic these stitched-together models are is the subject of debate. Free docking methods employ multi-step strategies with a broad initial search using the rigid unbound structures to identify approximate bound states, which are then flexibly refined. Any such strategy requires the rigidly docked state to be sufficiently close to the bound structure, which is often not true for large backbone changes. In Section 1.3.1, I discuss avenues for incorporating flexibility during the broad search, the implementation of which is detailed in Chapter 2. While I focus on broad local docking, parallels can be drawn for global docking.

In multi-component complexes (Figure 1.1D), the combinatorial explosion from having to simultaneously sample relative orientations of multiple bodies renders free docking impossible with current resources. Template-based modeling provides a promising approach in the rare cases where full complexes are crystallized.[50] Nevertheless, there is one class of complexes for which higher-order associations are readily tractable, *viz.* symmetric homomeric complexes. Symmetric homomers are ubiquitous across all domains of life, with studies estimating that they form 50-70% of all proteins.[51] A reduction in the degrees of freedom to

be sampled makes it possible to sample arbitrarily large symmetries. I discuss more about assembling symmetric complexes in Section 1.4 and Chapter 4.

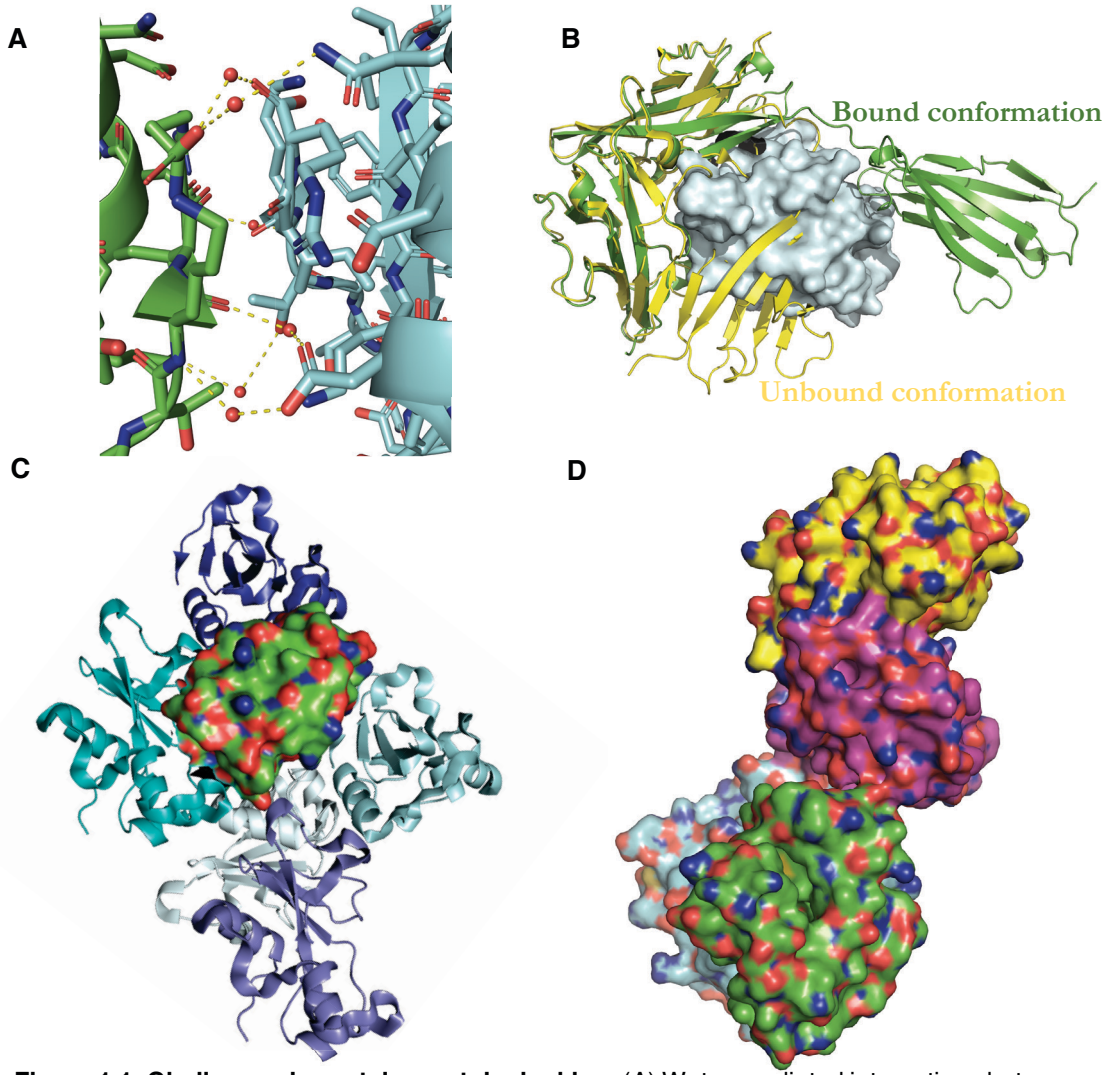


Figure 1.1: Challenges in protein–protein docking. (A) Water-mediated interactions between two proteins (green and cyan). Oxygens of the water molecules are shown as red spheres. (B) A large conformational change to recognize the partner (grey). The protein moves a domain from the unbound conformation (yellow) to the bound one (green) to accommodate the partner. (C) Low-energy binding modes of two proteins (green and blue). Both the correct orientation as well as the correct conformation of the flexible protein (hues of blue) needs to be determined. (D) A multi-component complex (green, cyan, pink and yellow).

1.3. Flexible-backbone protein docking

While docking together two rigid bodies, we sample along three translational and three rotational dimensions. For flexible proteins, however, the atoms within a monomer are free to move relative to each other. For a general case, this problem expands into a $3(n_1 + n_2)$ dimensional search problem, where n_1 and n_2 are the number of atoms in the two monomers, respectively. Fortunately, chemical bonds allow us to explore all feasible intra-monomer motions along mobile dihedral angles. In general, a protein molecule has $2(r - 1)$ mobile backbone dihedrals, where r is the number of residues. As a result, the dimensionality of a flexible free docking search is $6 + 2(r_1 - 1) + 2(r_2 - 1)$, where r_1 and r_2 are the number of residues in the two monomers, respectively. This value is $O(100)$ dimensions for typical proteins, which still makes it computationally infeasible to sample explicitly.[52]

Further, diversity in the kinds and magnitudes of motion limits our ability to generalize sampling strategies. Figure 1.2 shows three commonly observed motions between the unbound (yellow) and the bound (green) states when binding the partner (grey). Large concerted motions are observed throughout Ribonuclease inhibitor when binding Ribonuclease A (Figure 1.2A). Antibodies recognize the cognate antigens through pliable loops that complement the epitope surface (Figure 1.2B). A secondary-structure change involving the unfolding of a beta-hairpin in ADP-ribosylation factor 1 causes the it to bind Arno (Figure 1.2C). *A priori* information about the type of motion or the mobile elements is rare, but significantly reduces search space.

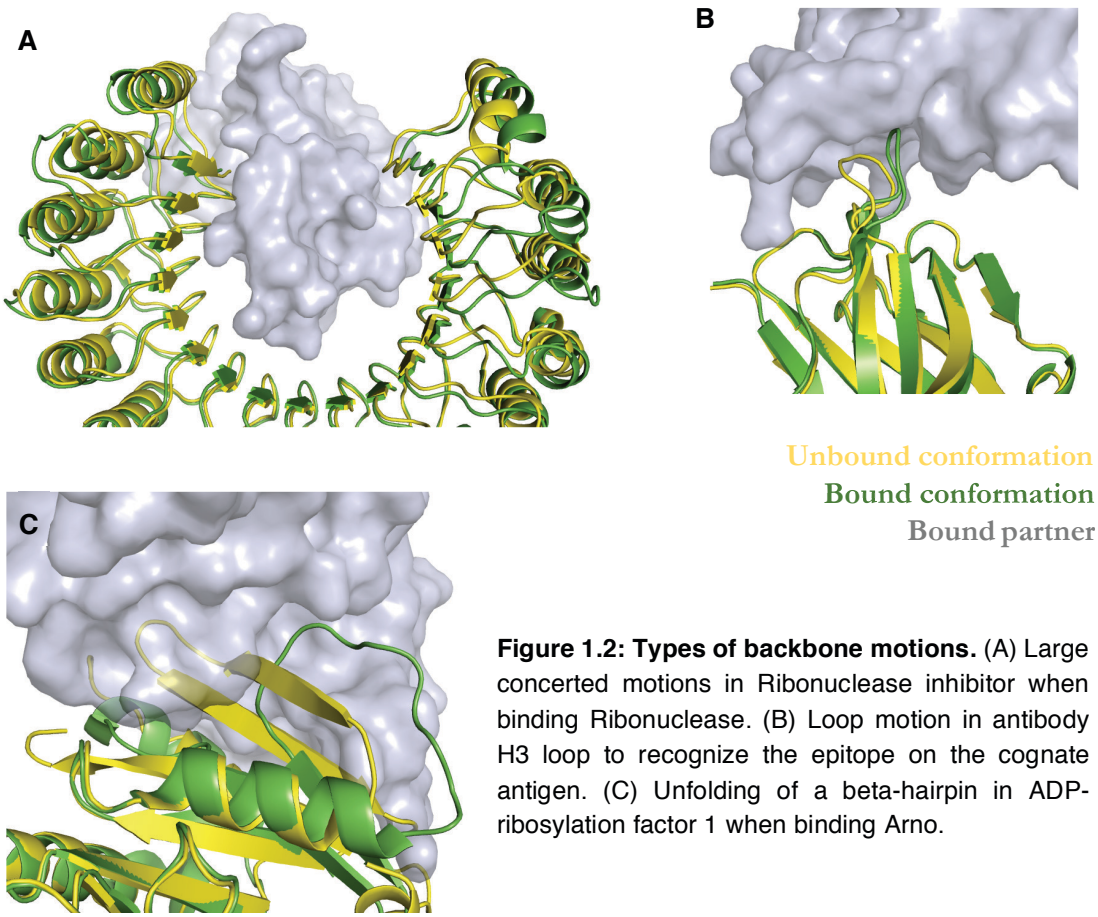


Figure 1.2: Types of backbone motions. (A) Large concerted motions in Ribonuclease inhibitor when binding Ribonuclease. (B) Loop motion in antibody H3 loop to recognize the epitope on the cognate antigen. (C) Unfolding of a beta-hairpin in ADP-ribosylation factor 1 when binding Arno.

Another source of flexibility is side-chain conformational change. Fortunately, pre-computed rotamer libraries have enabled rapid sampling of side-chain conformations.[53] Moreover, unlike backbone errors, errors in side-chain placement are not propagated through the model. Thus, unless the error is at a binding hotspot residue, the consequence on the overall docking model is limited. For the remainder of this dissertation, when referring to flexibility, I imply backbone flexibility, unless otherwise stated.

1.3.1. Simulating mechanisms of conformational change

Protein–protein association consists of two steps: first, electrostatically-guided diffusional formation of an initial encounter complex, and second, rearrangement of the partner proteins to form the fully-bound state.[54] For rigid-body assembly, docking algorithms exploit the concept of a 'lock and key' fit[55] in both a coarse search for the encounter complex and a finer search for the bound state. To capture flexibility, two kinetic mechanisms are frequently imitated: induced fit[56] and conformational selection,[57] which are not mutually exclusive.[58] The primary distinguishing feature between the two mechanisms is whether conformational change occurs before or after the formation of the encounter complex.

In induced fit, the partners in the encounter complex mutually adjust their shapes to enable the tightest pack. The simultaneous involvement of both the proteins makes this inherently difficult to simulate for large motions due to a high dimensionality of the search space. As a result, most induced-fit methods are limited to minor structural rearrangements like side-chain packing and interface dihedral motions.[26,59–61] However, if the encounter complex is recognized correctly, the steric constraints imposed by the partner can greatly reduce the overall search space. In Chapter 4, I demonstrate the efficacy of induced-fit refinement for large homomeric complexes. In Chapter 5, I propose a potential data-driven approach for identifying encounter complexes to enable large induced-fit motions.

Biomolecules exist in an equilibrium of conformational states, including in the monomeric form. During conformational selection, the partners shift the equilibrium from unbound-like

states to bound-like states. As the flexibility is inherent to each protein monomer, conformational selection is easier to simulate for large motions, thus reducing the computational complexity by many orders. Several perturbation algorithms exist to generate an ensemble of states from an unbound structure with varying degrees of overlap between the predicted and the actual unbound-to-bound conformational change.[62] Although a promising avenue to incorporate backbone change, docking using ensembles increases computational time (as compared to rigid docking) and leads to a plethora of false positives.[63] In Chapter 2, I develop a sampling algorithm that addresses both these issues and use it to dock moderately-flexible complexes.

1.4. Symmetrical homomeric proteins

The majority of proteins occur naturally as symmetric homomeric complexes,[51] making them an attractive target for modeling. Many of these proteins are transmembrane assemblies with important pharmaceutical applications, but they are beyond the scope of this discussion. For soluble proteins, the two most commonly observed point symmetry groups are cyclic symmetries, where the subunits are arranged around an axis of symmetry, and dihedral symmetries, where—in addition to an axis of rotation—there is a perpendicular two-fold axis. Evolutionary pressures like increased stability and finer functional control drive proteins towards forming larger assemblies.[64] Unfortunately, with increasing complex size, the resolution of structure determination methods suffers.

To predict atomic interactions in cyclic and dihedral complexes, symmetric docking protocols have been developed, covering all the aforementioned classes of methods.[65–69] In addition, for small proteins requiring oligomerization as part of the folding process, an approach has been developed that couples folding and association.[70] Common to all these methods is the idea of 'instant symmetrization', where all operations performed on one principal subunit are replicated on the other subunits such that they retain their symmetric relationship. While this massive dimensionality reduction enables us to sample multi-component complexes, it does not imitate natural association: the probability that a large number of monomers simultaneously orient in such a specific arrangement is infinitesimal, even with electrostatically-guided diffusion. Consequently, the binding energy landscape seen by such protocols has idiosyncrasies that require special attention. In Chapter 4, I investigate the shape of this landscape near the fully-bound native state.

In a recent study, four leading protocols were tested on a benchmark of 248 cyclic and dihedral complexes for their global docking accuracy.[69] Despite this benchmark being heavily biased towards the simplest, most frequently found,[71] and most easily modeled[72] symmetry, *viz.* C2, none of the methods could produce near-native models for majority of the complexes. In Chapter 4, I fill the gaps in knowledge by developing a next-generation symmetric docking protocol and testing it on a more balanced benchmark, which I also compile.

1.5. Rosetta

The Rosetta macromolecular modeling suite is a premier *in silico* framework for biomolecular structure prediction and design.[73] The underlying philosophy of Rosetta is Anfinsen's thermodynamic hypothesis, *i.e.* the observed native state of any protein is the unique global minimum free energy conformation.[74] To find the global minimum, most applications in Rosetta follow a Monte Carlo-plus-minimization (MCM) approach,[75] including RosettaDock[26] for docking hetero-dimers and SymDock[67] for docking symmetric homomers. In this approach, inter-monomer rigid-body transformations as well as intra-monomer backbone and side-chain motions are randomly sampled from a defined set of 'moves'. If the energy of the new state is lower than that of the old one, the move is accepted, else the Metropolis acceptance criterion is used, *i.e.* the probability of acceptance is sampled from a Boltzmann distribution: $P(\text{new state}) = e^{-\frac{E_2 - E_1}{kT}}$, where E_1 and E_2 are the respective energies of the old and the new states, k is the Boltzmann constant, and T is the temperature. After a series of attempted moves, the energy of the system is minimized by gradient descent along dihedral angles or atomic coordinates to arrive at a local minimum, and the process is repeated with the intention of finding a new local minimum. Although this approach does not simulate the natural dynamics of the system, with the right move set and energy function, it facilitates rapid sampling of multiple local minima to arrive at a near-global minimum.

Both RosettaDock and SymDock are multi-stage protocols, with an initial coarse-grained phase for broad searches. In the coarse-grained phase, the side chains are approximated by a

pseudo-atom called the centroid atom (Figure 1.3A). (Energy calculations in Rosetta scale as the square of the number of atoms; hence reducing the number of atoms speeds up calculations tremendously.) In the second all-atom phase, side chains are reintroduced and the encounter complex is refined to form the fully bound complex. To make better docking protocols in Rosetta, I had to consider two questions in both the phases. First, how can I choose a better move set (also known as the 'sampling problem')? Second, if near-native structures are indeed sampled, how can I discriminate them from non-native structures (also known as the 'scoring problem')?

1.5.1. Sampling in RosettaDock and SymDock

Rigid-body docking occurs through hundreds of random translational and rotational perturbations, starting with magnitudes of 0.5 Å and 7°, respectively. The magnitudes of these moves are modulated on the fly, depending on the fraction of perturbations accepted.[26] The assumption is that larger moves are more likely to be rejected as we approach the encounter complex state. As these moves are too small for broad sampling, each trajectory is started with a different initial orientation. For RosettaDock, in theory, global docking can be performed with a sufficient number of independent trajectories, but this number is too large for most proteins. In my opinion, the best use of RosettaDock is to perform a broad local search, where some information about the binding domain or region is known. For most proteins, approximately 5,000 independent trajectories with random perturbations of 3 Å and 8° to manually aligned monomers give enough initial orientations. In SymDock, as operations are

performed only on one protein, in most cases, global docking requires fewer than 5,000 independent trajectories. Moreover, the sheer depth of the binding funnel in homomers (discussed in detail in Chapter 4) guides the rigid-body search much more effectively. In the all-atom phase of both the protocols, smaller moves of 0.1 Å and 5° are performed to reduce rejection rates.

Conformational selection is simulated by pre-generating ensembles of backbone conformations for each monomer and attempting to identify the backbone pair that fits best. For this method to be successful, the ensemble generation method must produce a conformation where the interface backbone is close to the bound state. Most ensemble generation methods vary the mobile backbone torsions (φ and ψ) or the atomic positions without disrupting the bond lengths and bond angles. In RosettaDock, ensembles of both partners are docked simultaneously in the coarse-grained phase by repeatedly swapping the current backbone with another one from the respective pre-generated libraries. Backbone pairs with the best interface sterics for a given orientation are selected for refinement. In SymDock, independent trajectories are run for each backbone conformation in the ensemble.

Induced fit is simulated in the all-atom phase by repacking the side chains at the interface of the encounter complex. The side-chain conformations are chosen from a backbone-dependent rotamer library.[53] In Chapter 4, I systematically vary the backbone torsions for subunits in a symmetric homomer for larger induced fit motions.

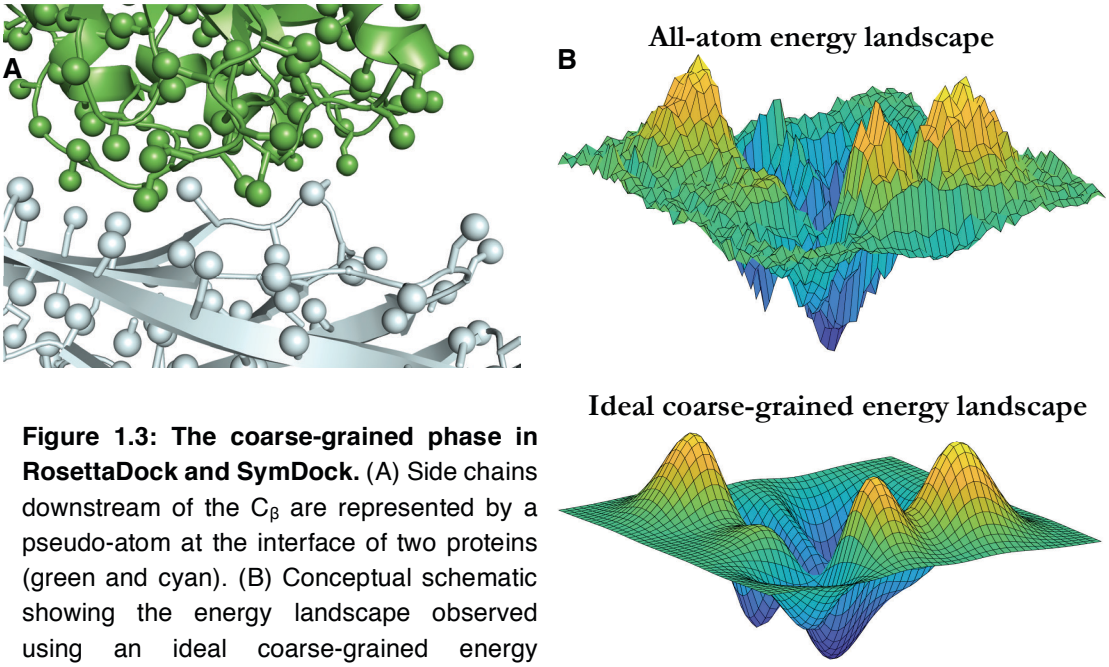


Figure 1.3: The coarse-grained phase in RosettaDock and SymDock. (A) Side chains downstream of the C_β are represented by a pseudo-atom at the interface of two proteins (green and cyan). (B) Conceptual schematic showing the energy landscape observed using an ideal coarse-grained energy function. It is less rugged than an all-atom landscape, but captures the broad features.

1.5.2. Scoring in RosettaDock and SymDock

In Rosetta, the energy of a state is estimated through mathematical functions built with the primary objective of swiftly discriminating native-like states. These functions score a state based on its atomic configuration using (a) physical laws governing electrostatics, van der Waals interactions, and solvation, (b) empirical observations on the geometry of hydrogen and disulfide bonds, and (c) statistical potentials describing torsional preferences, inter-chain contacts, interface residue orientations, and prevalence of residue types in a given chemical environment. To speed up calculations, score functions are linear combinations of one- and two-body terms, each of whose weights are optimized for a given application.

An ideal coarse-grained score function should broadly recover major peaks and troughs found in the all-atom energy landscape while reducing the overall ruggedness to facilitate a smooth search (Figure 1.3B). As this representation is missing side-chain atoms, any binding energy estimate at this level will fail to capture important interactions in the interface. As a result, the existing score function, called centroid score, often failed to discriminate near-native binding modes from spurious ones.[76] In Chapters 2 and 4, I test a novel coarse-grained scoring scheme optimized by my colleague, Dr. Nicholas Marze, based on the residue-pair transform framework developed by Dr. William Sheffler. Using just the relative orientations of the backbone atoms of interacting residues, this scheme uses an empirical potential to estimate the all-atom energy.

Many protocols in Rosetta have custom all-atom score functions, which have been optimized for specialized applications like docking. However, there also exists a general all-atom score function (called *REF2015*) parametrized for a large number of applications,[77] which was thoroughly reviewed recently.[78] In Chapters 2 and 4, I update the score function used in the all-atom phase of RosettaDock and SymDock to this all-purpose one. As they go hand in hand, I also optimize the sampling algorithm to best utilize this score function.

1.6. Outline of the dissertation

In the remainder of this dissertation, I describe the advances that I have made towards addressing the challenges of flexible-backbone docking and symmetric homomer docking. I validate my enhancements on test sets and blind prediction competitions.

In Chapter 2 (previously published[79]), I detail the development of RosettaDock 4.0, which combines a novel, six-dimensional coarse-grained score function (optimized by my colleague, Dr. Nicholas Marze) with efficient, adaptive conformational selection to sample libraries of hundreds of backbone conformations. RosettaDock 4.0 is the first method that successfully docks ~50% of flexible protein complexes with backbone conformational change of up to 2.2 Å. Further, I demonstrate that most failures are caused by the dearth of good ensemble generation methods, and not because of sampling or scoring inadequacies of the protocol.

In Chapter 3, I discuss my performance as part of the Gray laboratory group in rounds 37–45 of the community-wide blind prediction experiment called Critical Assessment of PRedicted Interactions (CAPRI). I detail the methodology used to predict the structures of homomeric protein complexes, heteromeric protein complexes, and oligosaccharide–protein complexes. Lastly, based on our prediction failures, I underline areas of improvement, especially for modeling homomers.

In Chapter 4, I describe the development of SymDock2, a homomer docking protocol that combines the score function optimized in Chapter 2 with induced-fit refinement. This protocol addresses the weaknesses identified in CAPRI to record a docking success rate of 77% for cyclic complexes, which is significantly higher than any competing method.

In Chapter 5, I summarize my contributions to the field of protein–protein docking and comment on the remaining challenges. I also lay a roadmap for data-driven strategies to better simulate conformational selection and induced fit.

Chapter 2

Efficient flexible backbone protein–protein docking for challenging targets

[Previously published as Marze, N. A.†, Roy Burman, S. S.†, Sheffler W., & Gray, J. J. Efficient flexible backbone protein–protein docking for challenging targets. *Bioinformatics* (2018); doi:10.1093/bioinformatics/bty355. †These authors contributed equally.

Reprinted with the permission of the publisher, Oxford University Press with minor revisions.]

2.1. Overview

Binding-induced conformational changes challenge current computational docking algorithms by exponentially increasing the conformational space to be explored. To restrict this search to relevant space, some computational docking algorithms exploit the inherent flexibility of the protein monomers to simulate conformational selection from pre-generated

ensembles. As the ensemble size expands with increased flexibility, these methods struggle with efficiency and high false positive rates. Here, I develop and benchmark RosettaDock 4.0, which efficiently samples large conformational ensembles of flexible proteins and docks them using a novel, six-dimensional, coarse-grained score function. A strong discriminative ability allows an eight-fold higher enrichment of near-native candidate structures in the coarse-grained phase compared to RosettaDock 3.2. It adaptively samples 100 conformations each of the ligand and the receptor backbone while increasing computational time by only 20–80%. In local docking of a benchmark set of 88 proteins of varying degrees of flexibility, the expected success rate (defined as cases with ≥50% chance of achieving 3 near-native structures in the 5 top-ranked ones) for blind predictions after resampling is 77% for rigid complexes, 49% for moderately flexible complexes, and 31% for highly flexible complexes. These success rates on flexible complexes are a substantial step forward from all existing methods. Additionally, for highly flexible proteins, I demonstrate that when a suitable conformer generation method exists, the method successfully docks the complex.

2.2. Introduction

Proteins bind each other in a highly specific and regulated manner. Often, a change in conformation from the unbound to the bound state forms the basis of the protein's specificity and function in its interaction.[80–83] Since the beginning of the field, [84] conformational changes in proteins induced by binding have confounded protein–protein docking algorithms by greatly increasing the degrees of freedom to be sampled. While rotamer libraries have alleviated

the sampling challenges for surface side chains,[53] backbone flexibility remains the principal challenge in protein-protein docking. Previous studies have found limited success by varying the backbone along a restricted set of coordinates[60,85,86] or interface residues[59,61] or by docking a small number of backbone conformations of the two partners.[87–90] The most recent rounds of the blind docking challenge, Critical Assessment of PRediction of Interactions (CAPRI), demonstrated that protein flexibility is still a community-wide weakness, with flexible target complexes eliciting no successful predictions from any method.[91,92]

Flexible-backbone docking, as well as other key remaining protein–protein docking challenges such as global docking and docking of large multi-domain complexes, demands more algorithmic complexity to explore a larger conformational search space than rigid-body docking of small proteins.[62] Coarse-graining is commonly used to model longer time-scales and larger systems in a rapid, yet meaningful manner.[93,94] Score functions designed to navigate this reduced space smoothen the energy landscape to avoid getting stuck in local minima. While allowing orders-of-magnitude more conformational sampling, coarse-grained models are limited by their accuracy and typically require high-resolution refinement.

The consensus on the kinetic mechanism of many conformational changes is that the protein monomers exist in an equilibrium of multiple conformations from which the preferred conformations are selected during an initial encounter with the binding partner, and subsequently, localized structural rearrangements stimulated by the partner tightens the binding.[95,96] The former mechanism is called conformational selection, and it lends itself to coarse-graining as the discrete conformations can be individually sampled. However, large

conformational ensembles of flexible proteins multiply the computational demand and increase the false positive rates. Previous studies have used experimental data to create a minimal ensemble that captures the observed flexibility,[97] or have selected optimal conformations from a large ensemble with *a priori* knowledge of the native orientation,[98] but these data are seldom available. Thus, it is desirable to have a coarse-grained method that efficiently samples a sizeable ensemble while distinguishing spurious interfaces from the native interface. Smaller changes caused by induced fit are less suitable to be modeled at this resolution, but are more amenable to full-atom modeling.

RosettaDock has been among the top-performing methods for computational protein–protein docking.[99–103] Combining coarse-grained conformational selection with full-atom induced fit, RosettaDock 3.2 achieved successful docking predictions on a majority of rigid complexes (58%) in the Docking Benchmark 3.0 set.[104] On the more flexible targets, however, RosettaDock (like other methods) performed poorly, only achieving a successful docking prediction on 29% of the moderately flexible complexes and 14% of the highly flexible complexes. The performance in CAPRI rounds since the last advances mimicked the benchmark performance.[105] For flexible docking, the current protocol relies on sampling a pre-generated ensemble of monomer backbone conformations,[88] but increasing the ensemble size beyond 20 conformers is computationally infeasible. Additionally, the "centroid" score function used to discriminate near-native conformations from incorrect ones is not sufficiently accurate in the coarse-grained phase, where the search is the broadest.[76]

In this study, I pursued two avenues to address these computational limitations. First, to improve sampling efficiency, I developed a fast and scalable backbone sampling algorithm, Adaptive Conformer Selection (ACS), that modulates the frequency of conformer selection for each partner depending on the size and diversity of the ensemble. Second, to improve scoring efficiency, I used a fast and accurate scoring method, Motif Dock Score (MDS), based on the residue-pair transform (RPX) score, which was recently developed to design hydrophobic symmetric protein interfaces.[7] RPX score evaluates residue pairs using the 6D transformation needed to superimpose the residues' N–C_α–C backbone atoms onto each other. In a single lookup, RPX score queries this transformation against a pre-tabulated database of aliphatic amino acid pairs and their corresponding geometries and full-atom Rosetta scores. The pair score and sequence of the best amino acid pair from the database are then assigned to the queried residue pair. My colleague, Dr. Nicholas A. Marze derived and optimized MDS from the RPX basis in the context of the RosettaDock protocol, expanding it to all twenty amino acids and selecting for enrichment of near-native candidate structures.

I tested RosettaDock 4.0, which contains both ACS and MDS enhancements, on a subset of Docking Benchmark 5.0[83] to evaluate the relative performance versus RosettaDock 3.2, and other commonly used docking protocols. The performance in both the full benchmark set and the three flexibility-based subsets (rigid, medium-flexible, and highly flexible) showed significant improvements, most notably among previously intractable flexible-backbone complexes.

2.3. Results

RosettaDock is a Monte Carlo-plus-minimization algorithm[75] consisting of a low-resolution stage, which simulates conformer selection during the formation of the encounter complex, followed by a high-resolution stage, which simulates induced fit in the bound complex.[26,88] To produce a variety of starting states for the different trajectories, the ligand (the smaller protein) is first randomly rotated and translated about the receptor (the larger protein). In the low-resolution stage, side chains are replaced by coarse-grained "pseudoatoms", allowing the ligand to efficiently sample the interface by rigid-body movements in a smoothened energy landscape. These rigid-body moves are coupled with backbone conformation swaps where the current backbone conformations of the ligand and the receptor are swapped with different ones from a pre-generated ensemble of conformations. In the high-resolution stage, the side chains are reintroduced to the putative encounter complex and those at the interface are packed for tight binding. There is minimal rigid-body motion in this second stage.

2.3.1. Adaptive Conformer Selection

The previous version of RosettaDock, version 3.2, was optimized to handle small ensembles and hence had a fixed number of conformation swaps. This choice led to reduced sampling of near-bound conformations as the ensembles grew larger. In RosettaDock 4.0, I alleviate this problem by modulating the number of conformer swaps depending on the swap acceptance rate of the previous cycle. If the acceptance rate of the conformer swaps is under

30%, the ensemble is presumed to be large and diverse, and hence the probability of the conformer swap is increased by 25%; conversely, if the acceptance rate is 30% or more, the probability is reduced by 25%. This adjustment helps prevent unnecessary backbone sampling for small ensembles and those with similar backbones while increasing backbone sampling for diverse ensembles by up to 477% over the course of 8 cycles. I call this backbone variation method Adaptive Conformer Selection (ACS). Figure 2.1A shows the variation in conformer sampling frequencies for an example case of the ClpA chaperone–Clp protease adapter complex (PDB: 1R6Q), where the unbound to bound deviation of the C_α atoms at the interface is 1.4 Å for the chaperone and 2.0 Å for the protease. In this case, the protocol adapts to enable more trials of the protease backbone conformer swaps, and to a lesser effect the chaperone too.

Previously, to determine which backbone was to be swapped in during conformer swapping, RosettaDock calculated the partition function of the entire ensemble of backbones superimposed along the protein-protein interface. The constraints of the interface, steric and otherwise, penalized conformations with backbone variations near the interface, creating a high probability for the existing backbone to be reselected during the conformer swap. In the case of superoxide dismutase (PDB: 1JK9), 36% of the backbone swaps were self-swaps (Figure 2.1B). Moreover, if there are n_1 conformations of the receptor and n_2 conformations of the ligand, the partition function calculation required $O(n_1 \bullet n_2)$ time, which meant that it required 10^3 times longer for ensembles with 100 conformations each than for ensembles with 1 receptor conformation and 10 ligand conformations.[88] I replaced this expensive partition

function calculation with random conformer swaps, speeding up the protocol by as much as 12-fold and reducing self-swapping to 8% (approximately the inverse of the size of the ensemble).

2.3.1.1. Efficiency of conformer selection

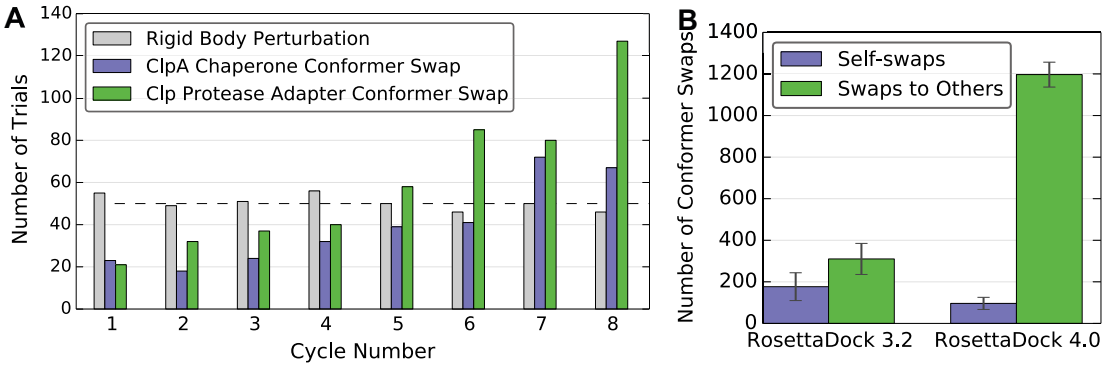


Figure 2.1. Amount of backbone sampling in RosettaDock 4.0. (A) Modulation of backbone conformer swap trials in Rosetta 4.0 for each of the first 8 cycles of Monte Carlo moves in the low-resolution search stage. The dashed line indicates the number of trials for each of the different moves in RosettaDock 3.2. Adaptive conformer selection in RosettaDock 4.0 ensures increased backbone swapping frequency for Clp protease adapter over ClpA chaperone, which is less flexible at the interface. (B) Comparison of the number of self-swaps versus swaps to other conformations in RosettaDock 3.2 versus Rosetta 4.0 for the highly flexible CCS metallochaperone:superoxide dismutase complex. RosettaDock 4.0 has increased backbone sampling both in the number and fraction of other conformations sampled.

ACS made RosettaDock 4.0 marginally faster than RosettaDock 3.2 for simulations with small ensembles of 1 receptor and 10 ligand conformations. The speed-up was pronounced

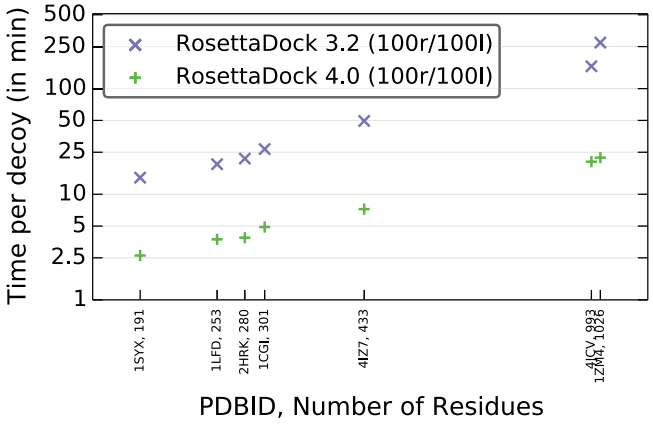


Figure 2.2. Time comparison of the docking protocols for large ensembles. Average time per decoy for RosettaDock 3.2 (x) and 4.0 (+) with ensembles having 100 receptor and 100 ligand conformations for complexes ranging from 191 to 1026 total residues. Adaptive Conformer Sampling makes RosettaDock 4.0 up to 12 times faster for cases with large interfaces.

when the ensembles of both partners have 100 conformations each. For protein complexes larger than 1000 total residues, for example, eEF2–ETA-bTAD complex (PDB: 1ZM4) with 204 residues in the ligand and 822 residues in the receptor, ACS was over 12 times faster than RosettaDock 3.2 (Figure 2.2). Thus, the ACS method scales up practically for larger ensembles.

2.3.2. Optimization and benchmarking of Motif Dock Score

For the recognition of the native interface during the broad, low-resolution search, docking requires a score function with predictive accuracy close to that of the well-tested full-atom score function. In earlier versions of RosettaDock, the low-resolution "centroid" score function relied on a single distance between potential interacting residues to score inter-chain contacts. This one-dimensional information was insufficient to represent the relative orientation of the two residues and consequently, their interaction. A statistical potential

derived by using two inter-residue distances (C_α–C_α and C_β–C_β) showed remarkable accuracy on Bcl-2 affinity predictions,[106] suggesting that with more information on relative orientation, it could be possible to distinguish native interfaces without representing the side chain in full. With this idea in mind, we developed Motif Dock Score (MDS) based on the residue-pair transform (RPX) framework[7] for interface design.

MDS calculates the 6-dimensional transform (3 rotations and 3 translations) needed to superimpose the backbone atoms of interacting residues, looks up the residue pair score from pre-generated tables, and sums scores over all such pairs. Each entry in these tables is the lowest full-atom score calculated for a pair of interface residues in the bin for the given relative backbone orientation. MDS depends on a discrete space tabulation of all-atom energies; therefore, Dr. Marze optimized the bin size of the scoring grid to 2 Å/22.5°. He also tested alternate underlying score functions to generate the residue pair motifs and recognized that the current Rosetta standard, *REF15*[77,78] had the highest average near-native enrichment of all score grids tested. Lastly, he added a van der Waals repulsive term to prevent protein partners from embedding in each other.

To evaluate the accuracy of local docking using MDS, its performance was compared against a baseline method, RosettaDock 3.2's centroid low-resolution docking mode, on a representative, nine-target benchmark set (set 2, section 2.5.2). For each of the two algorithms, 10,000 candidate structures were generated per complex. As examples, Figure 2.3 shows the Ras–RALGDS domain complex (PDBID: 1LFD) and BET3–TPC6 complex (PDBID: 2CFH) results. All candidate structures generated by the low-resolution phase of docking are

plotted, comparing their low-resolution score to their RMSD from the experimental bound structure. For the baseline score function (Figure 2.3A), the lowest-scoring models are nearly all incorrect with RMSD values from 7 to 22 Å, and few models under 6 Å are sampled at all. In contrast, with MDS (Figure 2.3B), a clear "funnel" can be seen in the plot, with the lowest-scoring models having low RMSD values from the native structure. The top-scoring structures are near-native indicating a successful discrimination. Further, if MDS was used to filter the candidate structures so that only the top 1 or 10% of low-resolution candidates were sent to the computationally intensive refinement stage, near-native structures would be included in the set. In contrast, filtering with the centroid score would eliminate the best structures. Docking results of BET3–TPC6 complex (Figures 2.3C–D) present a similar trend in that near-native models are lost when filtering on centroid score and can be retained by filtering on MDS. Appendix Table A.1 presents docking metrics for each of the nine complexes in the test set. Since this is a coarse-grained structure comparison, instead of the standard CAPRI metrics, near-native is defined as ligand $RMSD_{C\alpha} < 6$ Å. Significant improvements occur for all but the most flexible complexes.

To test whether MDS was unduly biased by existing structures of homologous interfaces in creating the score function, I removed all homologs of the proteins in Docking Benchmark 5.0 identified in the Dockground[19] and the PIFACE[18] libraries before building the motif tables. Appendix Table A.2 demonstrates that the performance of MDS with tables built after removal of the 8,126 homologs is similar to that with just the benchmark PDBs removed.

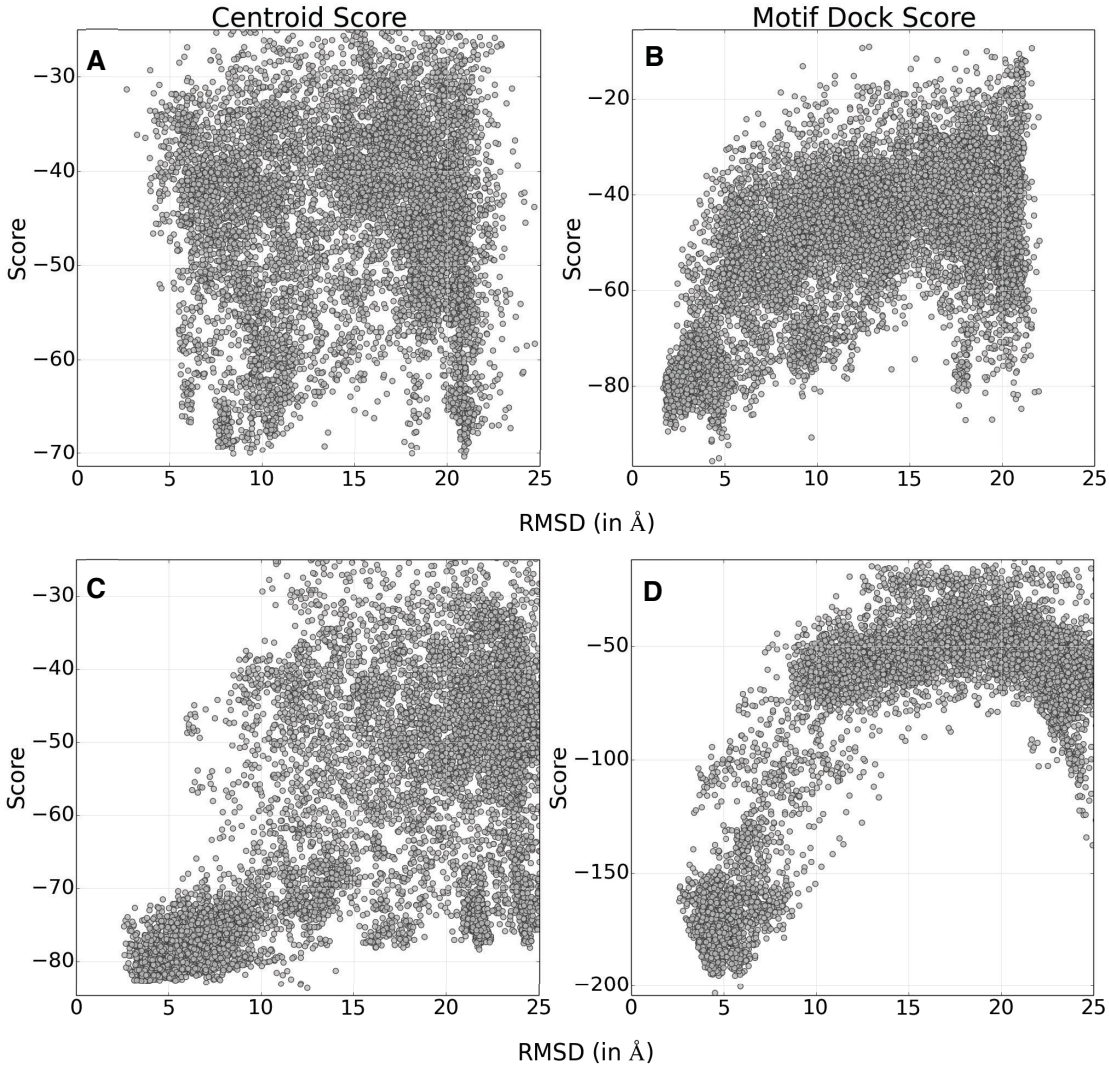


Figure 2.3: Low-resolution score vs. RMSD from native plots for two examples, viz. Ras:RALGDS domain complex (A and B) and BET3:TPC6 complex (C and D). (A and C) 10,000 models generated by RosettaDock 3.2 using the centroid score function, and (B and D) 10,000 models generated by RosettaDock 4.0 using motif dock score (MDS) function. (A) Centroid score does not generate many near-native candidate structures, and it cannot distinguish them from incorrect models. All metrics indicate failure: N5 = 0, N100 = 0, N1000 = 23. (B) MDS generates a large number of near-native candidate structures, and discriminates them from incorrect models. All metrics indicate success**:** N5 = 5, N100 = 95, N1000 = 750. (C) N5 = 1 indicates discrimination failure, but N100 = 86 and N1000 = 673 indicate that the broader set is enriched in near-native structures. (D) All metrics indicate success: N5 = 5, N100 = 98, N1000 = 813.

2.3.3. Advantage of using large and varied ensembles

In a blind prediction, where the location and extent of backbone motions is unknown, an ensemble generated using multiple conformation generation methods is more likely to contain a near-bound conformation than one generated from a single source. To delineate the gains made by using larger and more varied ensembles from the method improvements, I tested the two protocols, RosettaDock 3.2 and RosettaDock 4.0 with both small, similar ensembles and large, diverse ensembles. The small ensembles contained 1 receptor and 10 ligand conformations, all of which were made using Rosetta's Relax protocol.[107] The large ensembles contained a mixture of conformations generated using Relax, Backrub[108] and normal modes analysis (NMA)[109] for a total of 100 conformations of each docking partner.

The results of Ras–RALGDS domain complex (PDB: 1LFD) are depicted in Figure 2.4, where a large loop motion ($RMSD_{C\alpha}$ of 2.2 Å) helps the RALGDS domain interact with Ras. With RosettaDock 3.2 and the smaller ensembles (Figure 2.4A), few models could be classified as medium-quality, but, more critically, models with false interfaces had lower scores than these models, rendering the simulation unsuccessful on the *N5* metric. Docking with RosettaDock 4.0 and the smaller ensembles (Figure 2.4B) showed an increase in enrichment of medium-quality models and a successful dock. However, the RALGDS domain in the best scoring acceptable structures was rotated to enable the loop interact with Ras (Figure 2.4E). Using the larger ensembles, both RosettaDock 3.2 (Figure 2.4C) and 4.0 (Figure 2.4D) find deeper funnels due to the presence of conformations generated using the Backrub protocol where the contacting residues in the loop were within 0.2 Å of the bound structure. While

RosettaDock 3.2 recognizes these near-bound conformations, they are sampled infrequently. This distribution qualifies as a success on the *N5* metric, but the near-native enrichment was low. Moreover, it took an average of 19.3 minutes to generate a structure as opposed to 3.7 minutes for RosettaDock 4.0. With the protocol improvements in RosettaDock 4.0 and a large ensemble to sample, the best scoring models recovered up to 64% of the native residue-residue contacts (Figure 2.4F).

A similar phenomenon is observed in the case of glutamyl-tRNA synthetase–GU4 nucleic-binding protein 1 complex (PDB: 2HRK) where the interface undergoes a collective motion of 1.8 Å RMSD. A small, collective motion in the conformations generated by NMA prevents the backbone atoms of Asn-124 in glutamyl-tRNA synthetase from clashing with those of Arg-102 on GU4 nucleic-binding protein, allowing for a tighter interface (Appendix Figure A.9). These examples suggest that swift and adequate sampling of large ensembles generated from different sources better produces native-like interfaces.

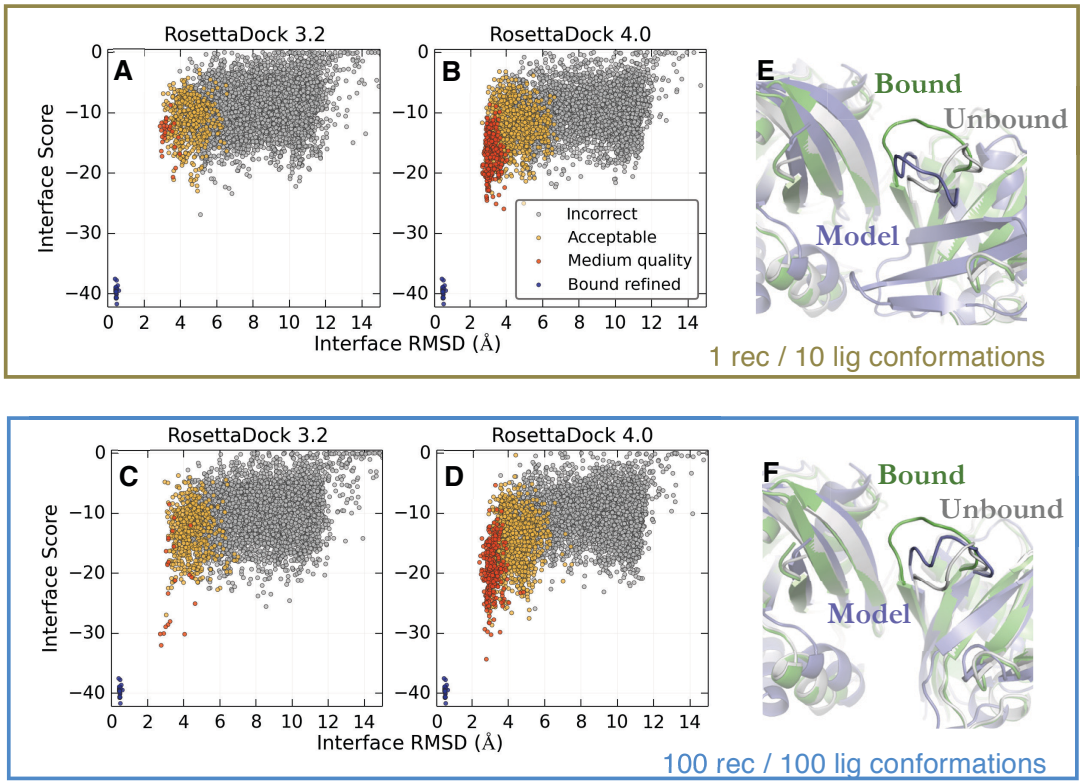


Figure 2.4. Comparison of docking protocols on Ras–RALGDS domain complex with different backbone ensembles. Interface score versus interface RMSD plots for docking simulation with (A) RosettaDock 3.2 and ensembles with 1 receptor and 10 ligand conformations generated by the Relax protocol, (B) RosettaDock 4.0 and ensembles with 1 receptor and 10 ligand conformations generated by the Relax protocol, (C) RosettaDock 3.2 and ensembles with 100 conformations each of the receptor and the ligand generated by the Relax, Backrub and NMA protocols, and (D) RosettaDock 4.0 and ensembles with 100 conformations each of the receptor and the ligand generated by the Relax, Backrub and NMA protocols. Colored points indicate CAPRI-quality category for each decoy, and the blue points provide a reference energy of the refined, bound crystal structure. (B) and (D) are enriched in medium-quality docked models as compared to (A) and (C), respectively. (C) has a deeper funnel than (A) owing to the inclusion of conformations generated by Backrub, which produces loop motions that mimic the unbound-bound conformational change. (D) has both a deep funnel and enhanced sampling. (E) The best docked structure (in blue) for runs with the smaller ensemble has the RALGDS domain rotated to find the interface interactions. (F) The best docked structure (in blue) for runs with the larger ensemble has a better overall RMSD and f_{nat} recovery. The superimposed unbound structures are in white and the bound structures are in green.

2.3.4. Evaluation of RosettaDock 4.0 on benchmark set

The ensemble generation methods used, *viz.* Rosetta Backrub, Rosetta Relax and NMA, have been shown to produce backbones that are between 1 and 4 Å RMSD from the unbound starting structure, with an average correlation of 0.4–0.5 to the experimentally determined displacements of the bound and unbound states.[62] The extent of motion suggested that the ensembles generated using these methods could be used to dock moderately flexible proteins. Thus, I built a benchmark set enriched with moderately flexible proteins to evaluate the RosettaDock 4.0 protocol.

I evaluated the accuracy of RosettaDock 4.0 for 43 complexes classified as medium-flexible, as well as for 32 classified as flexible and 13 classified as rigid, for a total of 88 targets (set 3, section 2.5.2). For each target, I pre-generated 100 conformations for both the ligand and the receptor ensembles. The three conformer generation methods produce motions in different directions and locations, and hence I increased the variability of the full ensembles by using 40 conformations made using NMA, 30 made using Backrub and 30 made using Relax. I then generated 5,000 local docked models using the full RosettaDock 4.0 protocol for each target. I also ran control simulations using the RosettaDock 3.2 protocol, also generating 5,000 candidate structures per target. For a fair comparison to the previously published accuracy metrics, I generated conformer ensembles for the control runs containing only 1 receptor conformation and 10 ligand conformations.

The ability of the two protocols to sample and discriminate near-native structures was evaluated using the bootstrapped *N5* average, ⟨*N5*⟩, both after the low-resolution stage and for the final models after the high-resolution stage. To evaluate the enrichment in the low-resolution stage alone, which dictates how many trajectories need to be run, I used the $\langle E_{1\%} \rangle$ metric. As summarized in Table 2.1, RosettaDock 4.0 shows significant performance gains over RosettaDock 3.2, particularly in the low-resolution phase. RosettaDock 4.0's near-native enrichment is improved markedly, with median $\langle E_{1\%} \rangle$ value of 2.5, implying that its very low-scoring sets are significantly enriched with near-native structures from the bulk candidate set. RosettaDock 3.2's median $\langle E_{1\%} \rangle$ value is 0.0, indicating that the very low-scoring set is devoid of near-native structures. Figure 2.5 compares enrichments of RosettaDock 3.2 versus RosettaDock 4.0 for each target. The $\langle E_{1\%} \rangle$ performance (Figure 2.5A) improves for 62 complexes in RosettaDock 4.0, most of which had zero enrichment previously. The performance is worse for 7 complexes, primarily due to favorable scoring of spurious interfaces. For the remaining 19 complexes, neither method was enriched in near-native decoys. RosettaDock 4.0 has an average low-resolution ⟨*N5*⟩ value of 1.3 across all targets, which implies that even after coarse-graining the side chains, more than one in the five top-scoring structures is near-native on average. This is approximately a ten-fold improvement over the corresponding average from RosettaDock 3.2. I defined the criterion for success discrimination is that the ⟨*N5*⟩ value should be 3 or higher. I see a seven-fold improvement in the number of expected low-resolution discrimination successes across the benchmark set (16.8 vs. 2.5 complexes). Pairwise target comparison (Figure 2.5B) shows that only 2 success

cases are lost from RosettaDock 3.2 to RosettaDock 4.0, while 13 additional successes are added.

While the low-resolution stage is improved using binned energy approximations, additional gains are possible in the high-resolution stage where all protein atoms are explicitly represented. After the full protocol with both low- and high-resolution stages, the average ⟨*N5*⟩ increases from 1.9 in RosettaDock 3.2, which represents a marginal failure, to 2.5 in RosettaDock 4.0, which represents the borderline for success. The expected number of successes in the benchmark set increases from 29.9 to 39.6 complexes, a 32% improvement. About half of the additional successes are gained from moderately-flexible complexes, with another quarter coming from flexible complexes, suggesting that RosettaDock 4.0 is better at capturing flexible backbones than RosettaDock 3.2. Additionally, although rigid complexes only comprise 15% of the benchmark set, they comprise 25% of the docking improvements, suggesting that in a more balanced benchmark set containing more rigid targets, the improvement in performance in RosettaDock 4.0 might be even larger. As shown in Figure 2.5C, while 23 complexes have full protocol ⟨*N5*⟩ values improved by 1 or more in the RosettaDock 4.0 simulations, 7 complexes have ⟨*N5*⟩ decreased by 1 or more. Detailed metrics and plots for each target can be found in Appendix Table A.3 and Figures A.1–A.6.

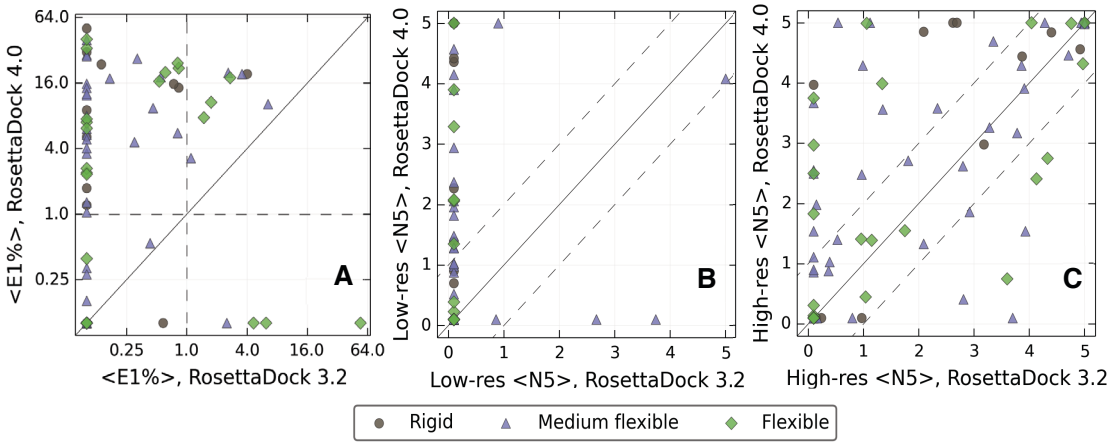


Figure 2.5. Comparison of performance metrics between RosettaDock 3.2 and RosettaDock 4.0 for individual complexes in the benchmark. Targets are represented by different symbols corresponding to their difficulty category (circle: rigid; triangle: medium; diamond: flexible). Points above the solid line represent better performance in RosettaDock 4.0, while points below the line represent better performance in RosettaDock 3.2. Comparison of (A) $\langle E_{1\%} \rangle$ enrichment values between the two protocols on a log-log axes. $\langle E_{1\%} \rangle$ shows marked improvement in the vast majority of the complexes. Dashed lines demarcate regions where the low-scoring set is enriched in near-native structures. Comparison of $\langle N5 \rangle$ values (B) after low-resolution stage, and (C) after high-resolution stage (full protocol). Dashed lines highlight the region in which the two protocols differ significantly, *i.e.* by more than one point in their $\langle N5 \rangle$ values. After the full protocol, 23 of the 88 complexes are modeled significantly better and 7 complexes are modeled significantly worse.

2.3.4.1. Ensembles doped with near-bound structures

We previously showed that when the RMSD gap between the closest conformation in the ensemble and the bound state exceeds 1 Å, induced fit methods are rarely able to access the binding funnel.[62] I observed similar results for the Docking Benchmark 5.0 difficult targets (cases with interface $RMSD_{C\alpha}$ > 2.2 Å). As none of the ensemble generation methods used move the backbone quite so far, neither RosettaDock 3.2 nor RosettaDock 4.0 performed

well on difficult targets. For example, the complex of SRP GTPase with FtsYh undergoes an interface conformational change of 2.67 Å RMSD (Appendix Figure A.11), and the docking run is only able to create a few acceptable predictions, but not rank them highly (Figure 2.6A). Both monomer backbones undergo about 3 Å of conformational change upon binding, but

Table 2.1. Summary of performance of RosettaDock 3.2 vs. RosettaDock 4.0 across an 88-target benchmark set. The ⟨*N5*⟩ values are the average bootstrapped *N5* values, both after the low-resolution stage and after the high-resolution stage (full protocol), with averages calculated across all targets in each flexibility category, as well as across the entire set. ⟨$E_{1\%}$⟩ is the median bootstrapped enrichment in the 1% top-scoring structures (after the low-resolution phase). Flexible target results include measurements with doped ensembles. The number of expected success cases, as calculated via bootstrapping is defined as follows: for *N5* values, ⟨*N5*⟩ ≥ 3; for ⟨$E_{1\%}$⟩, ⟨*N50*⟩ ≥ 15.

		RosettaDock 3.2			RosettaDock 4.0		
		Low-Res ⟨*N5*⟩	High-Res ⟨*N5*⟩	⟨$E_{1\%}$⟩	Low-Res ⟨*N5*⟩	High-Res ⟨*N5*⟩	⟨$E_{1\%}$⟩
Average Value	**Rigid Body**	0.0	2.7	0.0	2.2	3.5	9.0
	Medium	0.3	1.8	0.0	1.0	2.4	3.6
	Difficult	0.0	1.2	0.0	0.6	1.6	0.2
	Difficult (Doped)				0.7	2.2	2.9
	All	0.1	1.9	0.0	1.3	2.5	2.5
	All (Doped)				1.3	2.7	3.7
Expected Successes	**Rigid Body**	0.0	7.1	0.1	5.6	9.5	7.0
	Medium	2.5	15.4	2.8	7.6	20.2	13.0
	Difficult	0.0	7.4	0.3	3.6	9.8	5.0
	Difficult (Doped)				4.2	13.7	4.7
	All	2.5	29.9	3.2	16.8	39.6	25.1
	All (Doped)				17.4	43.4	24.8

the ensembles created from the unbound state do not contain any conformations closer than 2.5 Å from the bound state (Figure 2.6B).

For cases with large backbone variation, I wondered whether RosettaDock 4.0 could select a near-bound backbone if such structures were present in a large, diverse ensemble used in the conformer selection stage. Therefore, I tested docking using ensembles doped with near-native backbone structures. To generate near-bound structures, I used Rosetta's Relax protocol with pairwise C_α-C_α distance constraints to bias the simulation towards the known bound state (detailed in section 2.5.8). Using different constraint weights, I generated 10 conformers that were progressively nearer to the bound state, with the closest four conformations ranging from 0.59 to 0.81 Å RMSD from the bound structure for both receptor and ligand. To complete the ensemble, I mixed these 10 structures with an unbiased set of 36 NMA structures, 27 Backrub structures, and 27 Relax structures. For the SRP GTPase–FtsY complex (PDB: 2J7P), RosettaDock 4.0 produces structures using the full range of backbone conformations (Figure 2.6C) after the full protocol. Furthermore, the lowest-scoring docked structures are near-native (Figure 2.6B) and are chosen from the monomer backbones near the bound conformation (Figure 2.6D). Remarkably, even with just four near-bound backbones present in an ensemble of a hundred conformations with widely differing interface structures, RosettaDock 4.0 correctly recognizes these close conformations and docks them successfully. Figure 2.6D shows the correlation between closer backbones and better docked structures. Similar results are seen for others including the Pol III-ε–Hot complex (PDB: 2IDO), which has a 2.79 Å interface $RMSD_{C\alpha}$ between the unbound and bound states

(Appendix Figure A.10). In all, the doping method was able to add nearly 4 additional expected successes among the 32 difficult targets in the benchmark set. Detailed metrics for each target can be found in Appendix Table A.3 and Figures A.7–A.8.

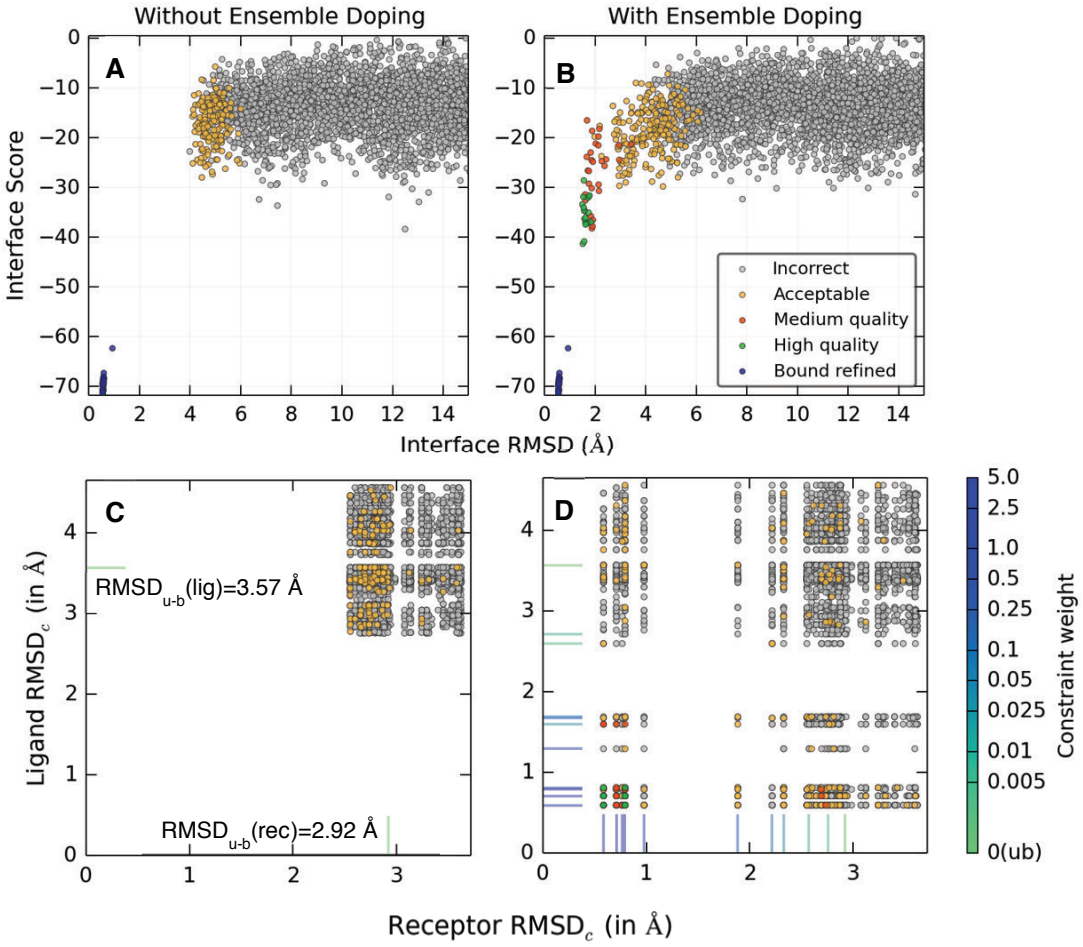


Figure 2.6. Improvement in docking performance of RosettaDock 4.0 by doping the ensemble with near-bound decoys for SRP GTPase–FtsY complex. Score versus RMSD plot of runs with (A) backbone conformations generated using NMA, Backrub and Relax protocols, and (B) ensembles doped with 10% near-bound conformations. (A) Without the ensemble doping, the simulations did not generate medium- or high-quality docked structures, and the acceptable structures did not score low enough to be discriminated from incorrect structures. (B) Ensemble doping generated deep docking funnels with high-quality structures. Colored points indicate CAPRI-quality category for each decoy, and the blue points provide a reference energy of the refined, bound crystal structure. (C and D) Plot of the contact-residue $RMSD_{C\alpha}$ from the bound conformation for the ligand and the receptor conformers selected after the docking simulation for (C) ensembles without near-native doping, and (D) ensembles with 10% near-bound conformations doped. The RMSD values of the unbound conformations are marked with a green line segment, and those of the near-bound conformations are marked in

colors corresponding to the biasing constraint weight. (C) The conformer generation methods are unable to generate sub-Å contact-residue $RMSD_{C\alpha}$ structures starting from the unbound ligand conformation (with $RMSD_{C\alpha}$ of 3.57 Å) and the unbound receptor conformation (with $RMSD_{C\alpha}$ of 2.92 Å). (D) Four of the biased conformations of the ligand and five of the receptor are within 1 Å $RMSD_{C\alpha}$ from the bound state. RosettaDock 4.0 is able to recognize these close conformations, find the native-like interface and successfully dock the complex.

2.3.4.2. Improved efficiency for large ensembles

One of the principal aims was to create a protocol that scales well with increasing ensemble sizes. Figure 2.7 shows run time across the benchmark set. In 77 of the 88 complexes tested, for ensembles containing 100 conformations each, RosettaDock 4.0 requires only 20-80% more time than RosettaDock 3.2 with just 1 receptor and 10 ligand conformations. Time per

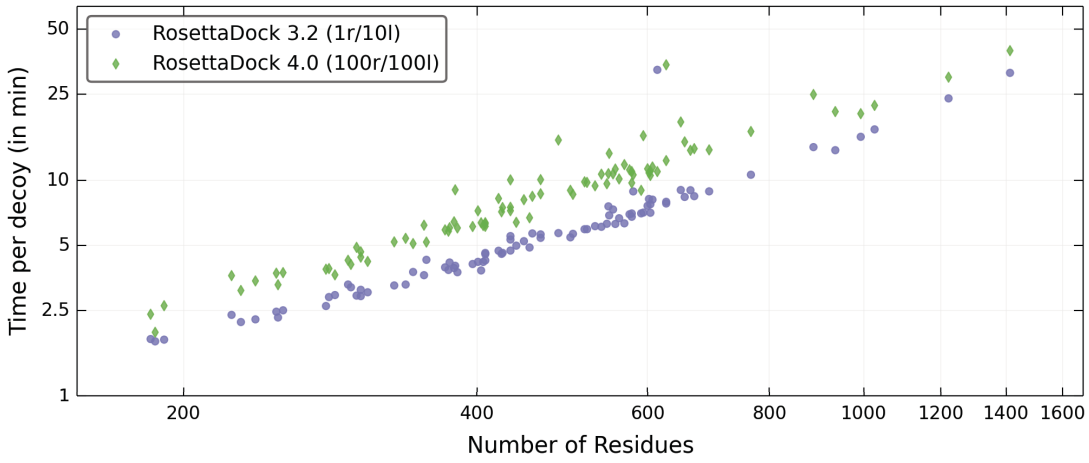


Figure 2.7. Efficiency of RosettaDock 4.0 on large ensembles. Despite sampling 100 conformations each of the receptor and the ligand as compared to 1 receptor and 10 ligand conformations in RosettaDock 3.2, the time per decoy for RosettaDock 4.0 is 20-80% more in 77 of the 88 targets tested.

structure scales as $\sim N_{res}^{1.4}$ for both RosettaDock 4.0 and RosettaDock 3.2, where N_{res} is the number of residues in the complex.

2.4. Discussion and conclusions

RosettaDock 4.0 combines two key advances. First, ACS now allows us to examine a variety of backbone motions introduced by different ensemble generation protocols. The protocol scales well with an increasing number of backbones by providing adequate sampling with a runtime overhead of merely 56% on average when testing 1000-times more backbone combinations. Second, the low-resolution scoring using MDS shows a marked improvement in accuracy over centroid scoring. MDS triples the number of targets in which the top 1% of models are significantly enriched with near-bound structures, and it is seven to nine times as effective for discriminating top models, as measured by the bootstrapped ⟨*N5*⟩ metric. More generally, MDS captures nearly all of the discriminatory power of the full-atom score function upon which it is based, exhibiting similar low-resolution and high-resolution *N5*, *N100*, and *N1000* metrics. Most importantly for a low-resolution score function, MDS achieves these gains in accuracy without sacrificing computational efficiency, running in roughly equivalent time to the centroid scoring method. It does require about 2 GB of additional memory to store the score table (requiring approximately 2.6 GB total compared to 0.6 GB for the baseline protocol). However, with modern computer architecture, this requirement is not prohibitive. With enhanced scoring and sampling, RosettaDock 4.0 can now select near-bound backbones in large, diverse ensembles for targets with significant changes at the interface.

RosettaDock 4.0 compares favorably to other docking protocols despite using more stringent success criteria. Table 2.2 summarizes recent published results from five leading docking methods: HADDOCK,[110] iATTRACT,[61] ClusPro,[111] ZDOCK,[112] and RosettaDock 3.2.[104] While the methods have different scopes and benchmarks, and report their results in different forms, we were able to assign an *N#* success metric (analogous to *N5*, *N100* etc.) to each method. In general, current methods are good at docking rigid-body targets (~50% accuracy or better), but they are all poor when the targets become more flexible (< ~30% accuracy on medium flexibility targets, <~15% on high flexibility targets). RosettaDock 4.0 maintains this level of accuracy for easy targets (77%) while showing dramatically improved accuracy for flexible targets, both among medium difficulty targets (49%) and high difficulty targets (31%). The performance of RosettaDock 4.0 on different success metrics is shown in Appendix Table A.1. To my knowledge, this is the first report of a protein docking protocol achieving ~50% accuracy on targets with backbone flexibility between 1 Å and 2 Å RMSD. Thus, RosettaDock 4.0 marks a key step toward a paradigm shift in protein–protein docking where complexes with backbone flexibility become tractable, which has long been a goal in the community.[92,113]

Table 2.2. Comparison of five leading docking methods with RosettaDock 4.0.

						Performance			
Method	Description	Flexibility?	Benchmark Set	Docking Search	Success Metric	All Targets	Rigid Targets	Medium Targets[e]	Flexible Targets[e]
HADDOCK (2017)	Restraint-based docking, minimization	Yes	CASP-CAPRI[f]	Mixed global/local	*N10*=1	16/25 (64%)	12/12 (100%)	4/13 (31%)	
ClusPro (2017)	FFT docking, cluster evaluation	No	CAPRI Rds. 13−35	Mixed global/local	*N10*=1	19/42 (45%)	12.5[b]/16 (78%)	6.5[b]/26 (25%)	
iATTRACT (2015)	Rigid-body docking, interface refinement	Yes	Docking Benchmark 4.0[g]	Global[a]	*N200*=30	64/166 (39%)	55/119 (46%)	9/28 (32%)	0/19 (0%)
ZDOCK (2011)	FFT docking, model evaluation	No	Docking Benchmark 4.0	Global	*N100*=1[c]	65/176 (37%)	58/121 (48%)	7/30 (23%)	0/25 (0%)
RosettaDock 3.2 (2011)	Monte Carlo docking, model evaluation	Yes	Docking Benchmark 4.0	Local	*N5*=3	56/115 (49%)	49/84 (58%)	5/17 (29%)	2/14 (14%)
RosettaDock 4.0 (this article)	Monte Carlo docking, model evaluation	Yes	Docking Benchmark 5.0[h]	Local	*N5*=3[d]	41/88 (47%)	10/13 (77%)	21/43 (49%)	10/32 (31%)

[a] Nearest-native structures from rigid-body docking selected for refinement. [b] Half successes awarded for targets with multiple binding sites evaluated, where at least one but not all binding sites are captured. [c] 2.5 Å cutoff for near-native structures. [d] Cases where bootstrapping gives ≥50% chance of N5≥3 are considered success-fully docked. [e] For CAPRI sets, medium & difficult targets are combined, comprising all targets without at least one high-quality prediction by any predictor. [f] Lensink et al., 2016.[91] [g] Hwang et al., 2010.[82] [h] Vreven et al., 2015.[83]

The limiting factor to successfully docking protein complexes with greater flexibility is now the ability to generate conformers within 0.7 Å of the bound state where MDS can start recognizing interfaces. Previously, our group compared seven commonly used methods to generate ensembles from monomers; while ensembles from most methods had ~50% directional overlap with the experimentally observed direction, the magnitudes of these motions were insufficient to reach the bound conformations.[62] Diversifying ensembles by pushing them along their top principal components may help close the gap. Another possible solution for proteins that have been crystalized in different contexts or have structurally diverse homologs is a distance geometry-based conformer selection method, which has recently been shown to span relevant conformational space.[114] Using energetic complementarity to the unbound partner as a means of generating and selecting conformers can also improve docking performance.[98]

While RosettaDock 4.0 makes large strides in conformer selection, the protocol still simulates induced fit only in the all-atom mode with small, rigid-body moves and side-chain packing at the interface. Other studies have shown significant contributions of induced fit, whether implemented via Cartesian minimization at the interface[61] or through contact-specific normal mode analysis.[115] Previous attempts to introduce flexibility at the interface in RosettaDock by varying backbone torsions resulted in 3-fold increased run times for the smallest targets.[116] Doing so by minimizing along Cartesian coordinates can slow the protocol down by more than 10 times (Section 5.2). These protocols were implemented in the high-resolution phase because the centroid score was not accurate for native discrimination. MDS

might now enable induced fit methods in the low-resolution phase, adding further backbone conformer sampling. Additionally, the accuracy of MDS means that low-resolution output structures might be filtered such that only a small fraction are sent to the expensive high-resolution phase. As such, MDS will be a critical component of the future ability to the RosettaDock protocol to induce a fit at the interface.

2.5. Methods

2.5.1. PDB curation

To create the score tables for motif dock score, I culled the Protein Data Bank [117] for all crystal structures containing two or more interacting protein chains and a resolution of 3.0 Å or better. I also removed any structures present in the Docking Benchmark 5.0 [83] to be used as a test set. I further removed all homologs of complexes in Docking Benchmark 5.0 and validated the lack of dependence on homologs. In the remaining set, PDB structures with more than two chains in their asymmetric unit were further divided such that one structure represented every pair of interacting protein chains in their asymmetric unit. The PDB structures were then stripped of all HETATM lines and non-canonical amino acids. My curated set contains 154,955 protein–protein complex structures from 103,017 PDB entries.

2.5.2. Benchmark set generation

Three benchmark sets using subsets of the Docking Benchmark 5.0 were built. The first, a set of eleven targets for rescoring, was randomly selected from the rigid-body subset of Docking Benchmark 5.0 to provide ample near-bound structures to optimize motif scoring's near-native discrimination ability. To generate the rescoring sets for each target, the standard RosettaDock protocol[104] was run on the unbound complex structures, including translation and rotation perturbations (3 Å translation, 8° rotation) to the ligand (the smaller protein partner) to disrupt existing interfaces. The second set, a small representative docking benchmark, was generated by selecting four rigid targets (1EFN, 1GLA, 2A1A, 2FJU), three medium-flexibility targets (1LFD, 2CFH, 3AAA), and two flexible targets of different categories (2OT3, 3F1P) from the Docking Benchmark 5.0. The third set, a larger representative docking benchmark, contained all targets in the second benchmark, as well as all rigid-body targets tested previously, [88] which still remained in Docking Benchmark 5.0 (13 in total), all remaining medium difficulty targets from Docking Benchmark 5.0, and 32 additional difficult targets chosen randomly from the Docking Benchmark 5.0 set. (I could not generate ensembles for the receptors of 1N2C, 3R9A, 1DE4 and 4GAM in reasonable time that would otherwise have been included in the benchmark set.)

2.5.3. Motif querying

Each of the 154,955 protein–protein complex structures in the protein interface set was loaded into Rosetta and scored with a full-atom score function; the resultant energies were

decomposed onto the set of interacting residue pairs. The system was queried for cross-chain pairs of residues with C_β atoms (C_α for glycine) within 10 Å of each other with a pair score below a constant energy cut-off (typically 0 kcal/mol; *i.e.* residue pairs that are net-attractive). For each residue pair in the filtered residue set, we calculated the six-dimensional transform needed to superimpose one amino acid backbone onto the other (three-dimensional Cartesian translation and three-dimensional Euler angle rotation). Each pair score was stored with its corresponding 6D-transform as a one-line motif.

2.5.4. Score grid generation

A score grid is initialized with a translational and rotational grid size. One by one, motifs are analyzed. The motif 6D-transform is binned, and the corresponding bin in the score grid is queried. If the bin is empty, the motif score is saved as the bin score. If the bin is populated, the old bin score and the motif score are compared, and the lower of the two is saved as the new bin score (see Supplementary Method S4 for further details).

2.5.5. Scoring with Motif Dock Score

RosettaDock 4.0 uses the same algorithmic framework as RosettaDock 3.2 described previously [26], with modernizations described in thereafter.[88,104,105] The standard low-resolution score function (`interchain_cen`) is replaced with a motif-based score function, called `motif_dock_score`. The score function consists of a new scoring term, motif_dock, and a clash penalty (`interchain_vdw`). The motif_dock term is a residue pair energy that acts

only on cross-chain residue pairs with C_α atoms within 10 Å of each other. The residue pairs are scored by calculating their 6D-transform, converting this to the hash value of the corresponding 6D bin, querying the hash table, and reporting the bin score. If the bin is empty (*i.e.* there are no matches for the hash), the pair score will either be zero if no penalty is used, or 0.5 kcal/mol, if a penalty is used.

2.5.6. Generation of backbone ensembles

To generate diversity in backbone conformations for the RosettaDock 4.0 runs, I used three conformer generation methods: perturbation of the backbones along the normal modes by 1 Å,[109] refinement using the Relax protocol in Rosetta,[107] and backbone flexing using the Rosetta Backrub protocol.[108] Since the normal mode analysis generated the largest deviations, I used 40 normal mode conformers, 30 Relax conformers and 30 Backrub conformers to comprise the ensemble of 100 conformers.

2.5.6.1. Relax

Rosetta FastRelax protocol is a refinement algorithm relying on iterations of side-chain packing and energy minimization in torsion space (φ, ψ and χ_i). Five cycles of refinement are carried out while ramping the repulsive part of the van der Waals score term. The command line was:

```
relax.linuxgccrelease
-in:file:s <PDB>  -nstruct 30    -relax:thorough
```

2.5.6.2. Normal mode analysis

Normal Mode Analysis is implemented as the `NormalModeRelaxMover` in Rosetta. I accessed this through an XML interface called RosettaScripts.[118] This protocol mixes motion along the first 5 normal modes, with perturbation of 1 Å. This is iterated with the `relax` protocol described previously. To prevent non-physical bond angles and bond lengths, I added a term to the score function to penalize deviations from ideal bond angles and lengths. The command line was:

```
rosettascripts.linuxgccrelease
-in:file:s <PDB>  -nstruct 40 -parser:protocol nma.xml
```

where `nma.xml` is:

```
<ROSETTASCRIPTS>
    <SCOREFXNS>
        <ScoreFunction name="ref_cart"
weights="ref2015_cart" />
    </SCOREFXNS>
    <RESIDUE_SELECTORS>
    </RESIDUE_SELECTORS>
    <TASKOPERATIONS>
    </TASKOPERATIONS>
    <FILTERS>
    </FILTERS>
    <MOVERS>
```

```
          <NormalModeRelax    name="nma"    cartesian="true"
centroid="false"        scorefxn="ref_cart"        nmodes="5"
mix_modes="true"    pertscale="1.0"    randomselect="false"
relaxmode="relax" nsample="20" cartesian_minimize="false" />
      </MOVERS>
      <APPLY_TO_POSE>
      </APPLY_TO_POSE>
      <PROTOCOLS>
          <Add mover="nma" />
      </PROTOCOLS>
      <OUTPUT scorefxn="ref_cart" />
  </ROSETTASCRIPTS>
```

2.5.6.3. Backrub

Rosetta Backrub protocol rotates segments of the protein backbone about an axis defined by the starting and ending atoms of the segment. This is followed by side chain packing. The command line was:

```
backrub.linuxgccrelease
     -in:file:s <PDB>  -nstruct 30    -backrub:ntrials
   20000     -backrub:mc_kt 0.6
```

2.5.7. Local docking simulations

Docking simulations were performed using two versions of RosettaDock, *viz.* 3.2[88] and 4.0 (developed in this article). The sampling and scoring enhancements implemented in

version 4.0 have been implemented in the low-resolution stage. The starting structure was generated by superimposing unbound monomers on the bound structure, moving them 15 Å apart, and rotating the smaller partner by 60° to scramble the interface. For each trajectory, a Gaussian random 3 Å and 8° perturbation provided different starting states. This allowed a broad local search. For motif dock score optimization and benchmarking runs, 10,000 and 5,000 decoys were generated per target, respectively.

2.5.7.1. Unbound-unbound simulations

Prior to docking simulations, the side chains of all backbone conformers, including the unbound state were optimized in isolation using the following command line:

```
docking_prepack_protocol.linuxgccrelease
-in:file:s   <PDB>     -nstruc 1
-ensemble1   <Receptor conformer list>
-ensemble2   <Ligand conformer list>
-partners    X1X2_X3    -detect_disulf true
-rebuild_disulf true   -ex1 -ex2aro
```

where $X_1X_2\_X_3$ are the ch**ain** ID's for the receptor (X_1 and X_2) and the ligand (X_3). Using this pre-packed structure, I then performed the docking simulations using the command line:

```
docking_protocol.linuxgccrelease
-in:file:s   <Pre-packed PDB>
-in:file:native   <Bound-state PDB>
```

(for calculation of metrics like f_{nat}, I_rmsd and L_rmsd)

```
-unboundrot  <PDB> -nstruct 5000 (or 10000)
```

```
-ensemble1   <Receptor conformer list>
-ensemble2   <Ligand conformer list>
-partners    X1X2_X3    -dock_pert 3 8 -spin
-detect_disulf true    -rebuild_disulf true
-ex1    -ex2aro
```

For some complexes, where the unbound and bound states had abnormally long disulfide bonds, I added:

```
-detect_disulf_tolerance 1.0 (or 2.0)
```

2.5.7.2. Low-resolution bound rescoring

In a similar experiment for the coarse-grained stage, I rescored using the command line:

```
docking_protocol.linuxgccrelease
-in:file:s <Pre-packed Bound-state PDB>
-in:file:native   <Bound-state PDB> (for calculation of metrics like f_nat, I_rmsd and L_rmsd)
-nstruct 100 -partners X1X2_X3    -dock_pert 0 0
-docking:low_res_protocol_only   -detect_disulf true
```

2.5.7.3. Low-resolution stage with Motif Dock Score

For all docking runs, to use Motif Dock Score and load the pre-tabulated values, I added:

```
-docking_low_res_score motif_dock_score
-mh:path:scores_BB_BB <Path to MDS tables>
-mh:score:use_ss1 false
```

```
-mh:score:use_ss2 false
-mh:score:use_aa1 true
-mh:score:use_aa2 true
```

2.5.8. Doping ensembles with near-bound structures

For highly flexible targets, the three ensemble generation methods often fail to generate near-bound conformations. Thus, I also evaluated the ability of RosettaDock 4.0 to discriminate bound-like structures with the help of conformations biased to resemble the bound state. Starting with the unbound conformation, I relaxed the structure while employing C_α distance constraints for all pairs of residues (except for adjacent residues). These C_α constraints are harmonic potentials with the mean distance set to the corresponding C_α–C_α distances in the bound conformation and a spring constant of 1 kcal/mol/Å^2. To create a library of intermediate structures, I set the weight of the constraints to 0.005, 0.01, 0.025, 0.05, 0.1, 0.2, 0.5, 1.0, 2.5, and 5.0. The ensembles were then doped with the 10 intermediate structures after proportionally removing 10 structures from the existing ensemble.

2.5.9. Near-native model criteria

To be counted as near-native, the high resolution models must meet the standard criteria for a CAPRI acceptable, medium-quality or high-quality model (*i.e.* have $f_{nat} \geq 0.1$ and, either ligand RMSD ≤ 10.0 Å or interface RMSD ≤ 4.0 Å) [119]. Here, f_{nat} is the ratio of the number of native residue-residue contacts in the predicted complex to the number of contacts in the experimental structure of the bound complex, ligand RMSD is the backbone RMSD of the

ligand molecule in the predicted complex versus the experimental complex upon receptor superposition, and interface RMSD is the interface heavy-atom RMSD after superposition of the backbone atoms of the interface residues. I use a more lenient measure for the low-resolution decoys (centroid RMSD ≤ 6.0 Å) to account for the limitations of measuring RMSD in the centroid phase (incompletely resolved side chains, lever-arm effects away from the interface etc.).

2.5.10. Benchmark evaluation and success metrics

I evaluate the results of the docking benchmark runs using two types of metrics: a top-scoring near-native model count (*N*#) and near-native enrichment values ($E_{N\%}$). I define *N*# as the number of near-native decoys among a set number (#) of top-scoring decoys after the high-resolution stage, analogous to the *N5* metric used in previous studies.[104] Docking runs with *N*# values above a given threshold are categorized as "successful". For *N5*, I define 3 near-native decoys as a success when evaluating docking protocols. I also use *N50*, *N100*, *N500*, and *N1000* (success thresholds of 15, 30, 75, and 150, respectively) to measure the sampling rates of near-native models in the top 1% and top 10% of models, respectively. Enrichment values are defined as:

$$E_{N\%} = \frac{\frac{\#\text{ near-native in top } N\%}{\#\text{ decoys in top } N\%}}{\frac{\#\text{ near-native}}{\#\text{ decoys}}}$$

I use $E_{1\%}$ and $E_{10\%}$ to measure the ability of the scoring schemes to enrich a model set. I calculate the expected value of *N*# and $E_{N\%}$ metrics by bootstrapping, i.e., resampling with

replacement from the available model set a number of models equal to the size of the set. This process was repeated 1,000 times, and bootstrapped averages are denoted by $\langle \cdot \rangle$.

Chapter 3

Predicting protein homomer, heteromer and oligosaccharide interactions using Rosetta in CAPRI rounds 37–45

[In CAPRI rounds 37–45, I led the Gray Laboratory team. This team also included Jeliazko R. Jeliazkov, Dr. Jason W. Labonte, Morgan L. Nance, Joseph H Lubin, Naireeta Biswas, and Dr. Jeffrey J. Gray. While describing individual targets, if I collaborated with any team members, I mention their names in the target header.]

3.1. Overview

The long-running blind prediction experiment, Critical Assessment of PRediction of Interactions (CAPRI) serves as a benchmark to assess the available macromolecular complex prediction methods.[39] We use this experiment to test our docking protocols and develop new functionalities. With every new round, the organizers of the challenge add to the complexity

of the modeling challenge by introducing larger complexes, new macromolecules, and multi-stage assemblies. From May 2016 to May 2018, as part of the Gray laboratory group, I participated in CAPRI rounds 37 through 45 during which we modeled 24 target complexes. The assessments of our prediction for 15 complexes were available at the time of writing, of which we successfully modeled 7. Additionally, we recognized and refined a near-native model generated by another group for an eighth successful prediction. RosettaDock 4.0, whose development I describe in the previous chapter, was available for use since round 39 (target 122) and allowed us to sample large ensembles of backbone conformations. Ten of the complexes had some degree of symmetry in the interactions and I used Rosetta SymDock to model them. In the process, I discovered some shortcomings of the protocol, which led to the development of the next-generation symmetric docking protocol, SymDock 2 described in the next chapter. In this chapter, I analyze each target in depth and based on our performance, I recognize areas for future development.

3.2. Introduction

With the explosion in genomic data availability and the ever-increasing accuracy of protein folding methods,[120] the ability to computationally model protein assemblies has taken center-stage. Protein docking methods provide a rapid way to model assemblies, and hence, their progress has been a key focus of computational biophysics. Over the years, various approaches have been developed, each with a different scope and ability to integrate experimental data, which I have briefly reviewed in the introductory chapter. Since 2001, a community-wide blind

experiment, Critical Assessment of PRediction of Interactions (CAPRI) has been used to assess the state-of-the-art in computational macromolecular docking.[39]

Every round of CAPRI has between one and ten target complexes that are thematically linked. For each target, the challenge consists of two phases: prediction and scoring. In the prediction phase, participants are challenged to model the structure of a biomolecular complex. Typically, only the sequences of the constituent proteins, stoichiometry of association, and in case of symmetric complexes, the point symmetry are provided. Participating groups have a limited time (usually three to six weeks) to prepare 10 alternative models. In addition to these 10, each group submits another 90 'decoy' models. In the scoring phase, the full set of 100 models from each group are then compiled by the organizers, anonymized and distributed to every participant. Participating groups then have five to seven days to refine and score the thousands of models from which they submit ten more predictions. The scoring phase allows teams to compensate for inadequacies in sampling during the prediction round and to test the discriminative ability of their scoring schemes. Finally, the experimentally determined structure, previously withheld from publication, is used to assess the submitted models. The organizers classify a model as high-quality, medium-quality, acceptable, or incorrect based on how it fares on three metrics: the fraction of native contacts recovered (f_{nat}), the root-mean-square-deviation of the backbone atoms from the native ligand after superimposing the bound receptor (Lrmsd), and the root-mean-square-deviation of the backbone atoms of the interface after superposition to the bound interface (Irmsd),.

Previous rounds of CAPRI necessitated the creation of protocols for flexible protein assembly and oligosaccharide-protein docking.[121,122] During rounds 37 through 45, we developed RosettaDock 4.0 to model flexible proteins (described in Chapter 2) and refined GlycanDock to predict oligosaccharide-protein interactions. Round 37 was a joint experiment between CAPRI and the Critical Assessment of Structure Prediction (CASP) in which preliminary monomer models submitted by CASP12 participants were provided to the CAPRI participants for docking.[123] This round comprised 11 targets ranging from homo-trimers to hetero-tetramers, 9 of which were symmetric homomers. Based on our performance on these homomers, I recognized the need for a new protocol utilizing the novel, coarse-grained score function we had optimized for RosettaDock 4.0, the development of which is described on the next chapter. Rounds 37, 39, and 40 challenged us with asymmetric multi-body assemblies, which required us to estimate the order of assembly. Rounds 39 and 42 required global docking while predicting the conformation of long, flexible loops. These targets gave us an opportunity to add functionality to dock single-chain camelid antibodies in our antibody docking protocol, SnugDock.[124] Due to time constraints, we could not participate in rounds 38 and 44.

In this chapter, I examine the biological relevance of each target complex, the challenge of modeling with the available information, the methodology we used, and if an assessment is available for our predictions, how we performed and what we could have done to improve prediction accuracy.

3.3. Methods and Results

We predicted the structures of 24 complexes in CAPRI rounds 37 through 45. Of the 15 that were evaluated at the time of writing, we achieved 1 high-quality, 2 medium-quality and 4 acceptable predictions. Most of the successfully modeled complexes were symmetric homomers with homologous structures available. Of the two successfully predicted heteromers, one required global docking with flexibility, which I consider to be one of the grand challenges of protein docking. In the scoring phase, we had one additional success based on a model submitted by another group for a target that also required global docking with flexibility. Tables 3.1–3.3 describe successful predictions, incorrect predictions, and targets to be evaluated, respectively.

Table 3.1: Summary of targets successfully modeled. The table lists the round, target number, name of the complex, the nature of the challenge, the methods used to model the complex, and the evaluation metrics for the best model that we submitted. The metrics are f_{nat}: the fraction of native contacts recovered, Lrmsd: root-mean-square-deviation of the backbone atoms from the native ligand after superimposing the receptor, Irmsd: root-mean-square-deviation of the backbone atoms of the interface after superposition to the bound structures, and quality: high-quality (***), medium-quality (**), acceptable (*), or incorrect (-) as evaluated by the CAPRI organizers.

					Best Model Evaluation			
Round	**Target**	**Complex Name**	**Challenge**	**Method(s) Used**	**f_{nat}**	**Lrmsd (Å)**	**Irmsd (Å)**	**Quality**
37	110	Fibre head domain of raptor adenovirus 1	Symmetric docking based on homolog	SymDock local refinement	0.67	2.35	1.73	**
	111	Fibre head domain of lizard adenovirus 2	Symmetric docking based on homolog	SymDock local refinement	0.75	0.98	0.73	***
	112	Fibre head domain of goose adenovirus 4	Symmetric docking based on homolog	SymDock local refinement	0.38	5.81	2.90	*
	118	Fructose bisphosphatase	Symmetric docking based on homolog	SymDock local refinement	0.41	1.72	1.07	**
	119	Alcohol dehydrogenase	pH-dependent symmetric docking based on homolog	SymDock local refinement, pHDock	0.61	9.92	2.99	*
	120	Group 1 dockerin–cohesin complex	Local docking based on homomer	RosettaDock with ensembles	0.25	4.96	3.80	*
39	122	IL-23–receptor complex	Multi-body docking, global docking	RosettaDock with ensembles	0.22	16.86	0.70	*

Table 3.2: Summary of targets modeled incorrectly. 'Best' corresponds to the model evaluation of the best model submitted by all CAPRI participants. The description of the other headers is the same as in Table 3.1.

					Best Model Evaluation			Best in CAPRI
Round	**Target**	**Complex Name**	**Challenge**	**Method(s) Used**	**f_{nat}**	**Lrmsd (Å)**	**Irmsd (Å)**	
37	113	CDI toxin–immunity complex	Flexible loop, global docking	ClusPro, RosettaDock with ensembles	0.06	18.33	11.43	*
	114	Ljungan virus protein	Low-quality monomer, symm. global docking	SymDock with ensembles	0.01	47.75	10.83	-
	116	Bifunctional histidine kinase	Variable domain linker, symmetric global docking	SymDock with ensembles	0.09	36.02	11.86	-
	117	Pins–Insc complex	Partial monomer unfolding, multi-body docking	RosettaDock with ensembles, SymDock	0.00	80.39	36.13	-
39	123	$PorM_{N\text{-}term}$–nb130 complex	Loop flexibility, global docking	RosettaAntibody, SnugDock, ClusPro	0.16	18.53	5.06	*
	124	$PorM_{C\text{-}term}$–nb02 complex	Loop flexibility, global docking	RosettaAntibody, SnugDock, ClusPro	0.00	27.42	14.15	-
42	131	CEACAM1–HopQ1 complex	Loop flexibility, global docking	CCD + fragments, RosettaDock with ensembles	0.04	39.28	11.61	**
	132[a]	CEACAM1–HopQ2 complex	Loop flexibility, global docking	CCD + fragments, RosettaDock with ensembles	0.00	34.79	10.42	**

[a] Refinement and scoring of the models of other participants resulted in a structure classified as acceptable.

Table 3.3: Summary of targets whose results are yet to be released. The header description is the same as in Table 3.1, but without best model evaluation metrics.

Round	Target	Complex Name	Challenge	Method(s) Used
37	115	GP1 domain of whitewater arroyo virus	Symmetric global docking	SymDock
40	125	NKR-P1–LLT1 complex	Multi-body docking, symmetric global docking	ClusPro, RosettaDock with ensembles, SymDock
41	126	1,5-α-L-arabinohexose bound to AbnE	Oligosaccharide docking	GlycanDock
	127	1,5-α-L-arabinopentose bound to AbnE	Oligosaccharide docking	GlycanDock
	128	1,5-α-L-arabinotetrose bound to AbnE	Oligosaccharide docking	GlycanDock
	129	1,5-α-L-arabinotriose bound to AbnE	Oligosaccharide docking	GlycanDock
	130	1,5-α-L-arabinopentose bound to $AbnB_{E201A}$	Oligosaccharide docking	GlycanDock
43	133	Designed colicin E2–Im2 complex	Loop modeling, docking based on homolog	KIC, RosettaDock with ensembles
45	136	Lysine decarboxylase	Symmetric docking based on homolog, complex size	Constrained relax, SymDock local refinement

3.3.1. Successes

3.3.1.1. Targets 110–112: Viral fibre head domains

The first three targets of round 37 were homo-trimeric fibre head domains from different viruses. Fibres are rod-like appendages that viruses like adenoviruses use to attach to the host cell. The trimeric fibre protein ends in a globular head domain involved in receptor binding.[125] Target 110 (T110) was the fibre head domain of raptor adenovirus 1, T111 was that of lizard adenovirus 2, and T112 was that of goose adenovirus 4, and the task was to predict the trimer quaternary structure.

Before we started docking the models, the organizers provided us with the initial monomer models submitted by CASP12 participants. First, I relaxed all the models using Rosetta FastRelax,[107] clustered those with similar backbone conformations, and chose between one and four monomer backbones to start docking. While I had monomer models for all three target complexes, I sought homologous complexes with similar oligomeric states. For T110, all homologous proteins had less than 40% sequence coverage and identity. Importantly, all these distant homologs lacked a predicted beta-hairpin (residues 358 to 373) present in T110, which I presumed to be essential for docking. T111 had a homolog with 94% sequence coverage and 50% identity, which strongly suggested that the native structure would resemble snake adenovirus 1 fibre head (PDB ID: 4D0V). For T112, the avian adenovirus CELO fibre head (2IUM) came the closest with 59% coverage and 27% identity.

For T110, due to the aforementioned beta-hairpin, arranging the subunits based on the homolog symmetry led to major atomic clashes, and I considered it to be an unsuitable initial placement. Instead, I performed independent symmetric global docking simulations with each of the aforementioned monomer conformations using Rosetta SymDock. As I was unsure of the beta-hairpin conformation, I also docked monomers where this region was truncated. For initial subunit placement for T111 and T112, I used subunit arrangements derived from their respective homologs. All three complexes were then docked using Rosetta SymDock to generate between 10,000 and 50,000 models per monomer.

For T110, the experimental crystal structure (5FJL)[126] is shown in in Figure 3.1A (grey). The native structure did indeed fold into a beta-hairpin in the predicted region, which is highlighted in red. On the superposed complex, the RMSD of the predicted beta-hairpin was 1.4 Å from the native. Our best model (yellow) recovered 67% of the native contacts across the subunits and had an Lrmsd of 2.35 Å from the native, which was the lowest amongst all models submitted by any group. While broadly correct, the model is not quite as compact as the crystal structure: the distance between the center of masses of the subunits is 24.2 Å for the crystal structures and 25.0 Å for the model.

The presence of a close homolog made T111 an easy target with multiple groups, including ourselves, predicting high-quality models. Our best model has a sub-angstrom Lrmsd from the native with 75% of the native contacts correctly predicted. The crystal structure of the target is still unreleased, but I assume it to be similar to snake adenovirus 1 fibre head (4D0V).

In Figure 3.1B, our best model is shown in orange superimposed on the grey crystal structure of the homolog.

Conversely, the absence of close homologs made T112 particularly hard to model both in the monomeric state and in the trimeric state. No group achieved a medium- or high-quality prediction. The Lrmsd and Irmsd of our best model was 5.8 Å and 2.9 Å, respectively and hence, it was classified as acceptable. A different model of ours was able to capture 53% of the native contacts, but had worse RMSD values. The crystal structure of the target is still unreleased and no close homolog is present for a visual comparison.

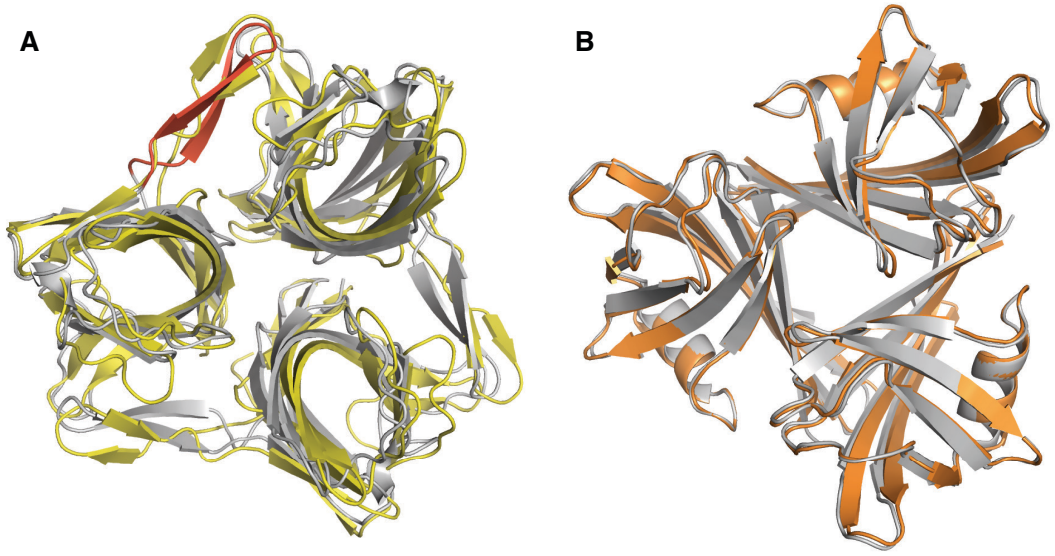


Figure 3.1: (A) **T110:** our best model of the fibre head domain of raptor adenovirus 1 (yellow) superimposed on the crystal structure (grey). Predicting and modeling the beta-hairpin in the native structure (red) was crucial to prediction success. (B) **T111:** our high-quality model of the fibre head domain of lizard adenovirus 2 (orange) superimposed on the crystal structure of a close homolog (grey), snake adenovirus 1 fibre head.

3.3.1.2. Target 118: Fructose bisphosphatase homo-octamer

T118 was a refinement challenge involving fructose 1,6-bisphosphatase from *Thermus thermophiles*. This enzyme converts fructose 1,6-bisphosphate to fructose 6-phosphate in the Calvin cycle. Although the organism is a hyperthermophile, we were not provided any temperature information about this target. I found a close homolog in fructose 1,6-bisphosphatase from a thermo-acidophilic archaeon, *Sulfolobus tokodaii* with 100% sequence coverage, 46% identity and the same D4 symmetry (3R1M).

As the structure of a close homolog was available, many CASP12 participants submitted convergent monomer models. I extracted symmetry information from the aforementioned homolog, arranged the monomer models and refined the complex using fixed-backbone

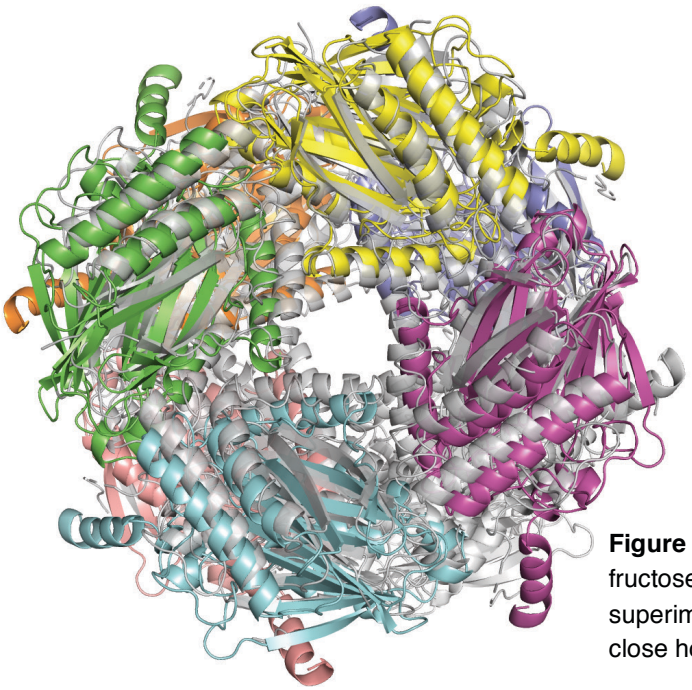


Figure 3.2: T118: our best model of fructose 1,6-bisphosphatase (color) superimposed on the crystal structure of a close homolog (grey).

refinement of SymDock. Figure 3.2 shows our best model in color and the crystal structure of the homolog in grey. The crystal structure of the target is yet to be released. The model recovered 41% of the native contacts with Lrmsd and Irmsd values of 1.7 Å and 1.0 Å, respectively, and hence, was classified as a medium-quality model. Another notable feature of this model was the increased inter-subunit distance of 43.3 Å compared to 39.0 Å in the homologous protein, which I presume to be similar to the native target. The best structure across all groups was classified as high-quality with 66% of the native contacts being recovered, but had 30 inter-chain clashes.

3.3.1.3. Target 119: Archaeal halo-thermophilic alcohol dehydrogenase

T119 challenged us with atypical modeling conditions. The homo-dimeric protein, alcohol dehydrogenase was from a halo-thermophilic archaeon expressed in a halo-mesophilic expression system. The behavior of this enzyme is pH dependent: in the pH range of 9.6–10.2, its oxidative reaction peaks, whereas at pH of 6.4, reduction reaction is dominant.[127] We were asked to predict the structure of the complex at pH 10.

First, we relaxed and selected monomer models from CASP12 participants. The closest homologous homo-dimer that I found was alcohol dehydrogenase 2 from the bacteria *Zymomonas mobilis* (3OWO). This homolog had 30% identity covering 98% of the sequence. Although the homologous protein was not from a halophile and was not expressed in pH 10, it still served as a starting stage. The two subunits of the homolog had extensive cross-beta sheet interactions along the N-termini. The N-termini interaction served as a hinge, where a

small error in the backbone would result in a drastically different rigid-body conformation. Unfortunately, this region of the target protein was predicted to be disordered and was quite different in all monomer models. As a result, I had to partly truncate the N-terminus.

I followed a two-pronged approach to model this target: on the one hand, I explored the homo-dimeric conformational space using SymDock; on the other hand, I sampled different protonation states at pH 10 with Rosetta pHDock[128]. I produced 10,000 docking models with each method and chose the most symmetric proteins interacting at the N-terminus for pHDock.

Figure 3.3 shows our best model in yellow, the crystal structure in grey, and the N-terminus of the crystal structure in red. Despite missing key interactions at the N-terminus, I was able to predict the rough placement of the subunits correctly, and hence our best model was adjudged acceptable. The model predicted 61% of the native contacts with Lrmsd and Irmsd

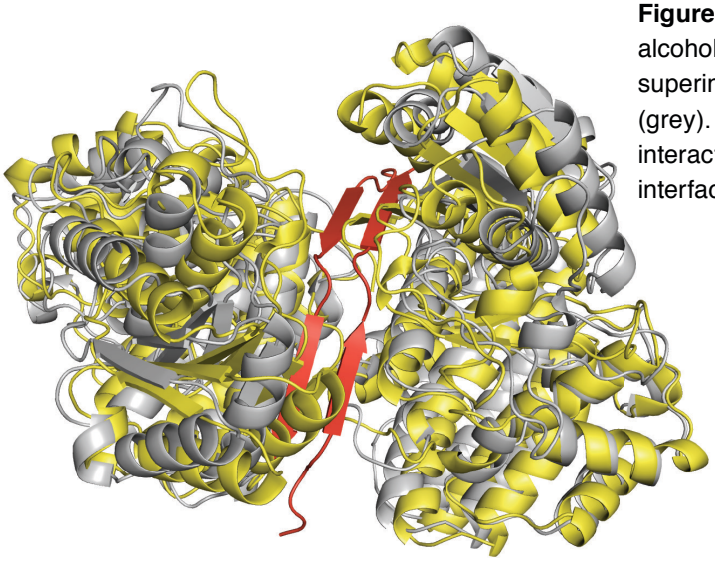


Figure 3.3: T119: our best model of alcohol dehydrogenase dimer (yellow) superimposed on the crystal structure (grey). The model is missing cross-beta interactions between the subunits at the interface (red).

values of 9.9 Å and 3.0 Å, respectively. This large difference of RMSD values arises from the aforementioned hinge motion, where a small change in the N-terminus backbone leads to large changes globally. The best model across all groups was a medium-quality model with Lrmsd and Irmsd values of 4.3 Å and 2.0 Å, respectively, but 81 inter-chain clashes.

3.3.1.4. Target 120: Group 1 dockerin–cohesin complex (with Joseph H. Lubin)

In anaerobic bacteria, a multi-enzyme complex called the cellulosome digests plant fibers. The assembly of this complex involves the binding of different enzyme-borne dockerin proteins (Doc) to cohesin modules of the non-enzymatic protein, scaffoldin (Sca). As different groups of dockerins have significantly different cohesin-binding interfaces, they have different binding modes for every cohesin.[129] Moreover, within the same species, each dockerin binds cohesins promiscuously with different binding modes. To compound the challenge of identifying the correct binding mode, their plasticity can be attributed to single residue changes.[130] T120 was a hetero-dimer of ScaB3 cohesin with Doc1a from *Ruminococcus flavefaciens*.

We started with homology models from CASP12 participants and relaxed them. Next, we searched for homologous complexes to create an initial placement of the monomers. A complex of group I dockerin and ScaB from *Acetivibrio cellulolyticus* (4UYQ) provided a starting point despite low homology with the individual proteins—the cohesin had 24% sequence coverage with 33% identity and the dockerin had 79% coverage with 37% identity. Starting

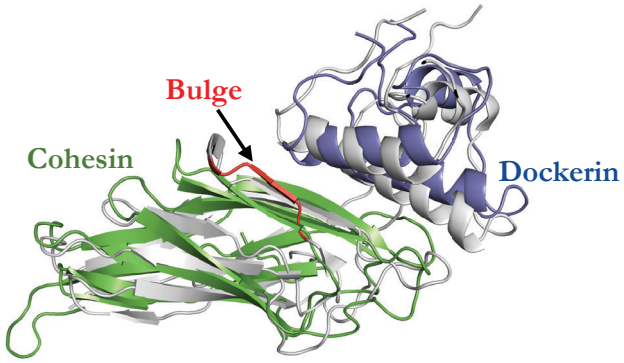


Figure 3.4: T120: our best model of group I dockerin (blue) - cohesion (green) complex superimposed on the crystal structure (grey). A bulge in the cohesion (red) was not modeled correctly leading to a small error in the rigid body orientation of the dockerin.

from an initial structure where the monomers were aligned to the homologous complex, we docked the target. While docking the proteins, we used an ensemble of 10 relaxed monomer models for each partner to explore alternate backbone conformations.

Figure 3.4 shows our best model in green (cohesin) and blue (dockerin) against the crystal structure in grey. This model was adjudged to be acceptable; no other group submitted a higher-ranking model. The bulge in the crystal structure of cohesin (highlighted in red) was not present in any of the homology-modeled cohesins. This bulge changed the rigid-body conformation of the dockerin and resulted in the dockerin having an Lrmsd of 4.9 Å. We correctly predicted 25% of the native contacts with an Irmsd of 3.8 Å.

3.3.1.5. Target 122: Human IL-23–receptor complex

Interleukin 23 (IL-23) is a pro-inflammatory cytokine, which plays a crucial role in the development of helper T-cells. IL-23 and its receptor, IL-23R participate in positive feedback

loop enhancing each other's expression.[131] For T122, we were asked to model this crucial interaction. IL-23 belongs to a unique family of hetero-dimeric cytokines, and hence, its interaction with its receptor is a multi-body docking problem. To further complicate docking, it binds two closely-related receptor chains, IL-12Rβ1 and IL-23R at two different sites. In this challenge, we were only required to model binding to the three extracellular domains of IL-23R.

Several crystal structures of IL-23 were available in the Protein Data Bank.[15] A disulfide bond held together its two subunits, IL-23A and IL-23B and hence, I expected their bound state to remain largely unchanged. I modeled the receptor, IL-23R using Modeller[132] based on multiple sequence alignment of homologs with manual input on the alignment of loop regions. In addition, I also used models from Robetta,[133] which used a different homolog as its template. From the variety of models obtained, it was apparent that the receptor might have inter-domain flexibility between its three domains. This flexibility rules out the possibility of global docking. A literature survey revealed that the binding site observed in other cytokine/cytokine receptor complexes in this family was likely used to bind IL-12Rβ1, which was not the receptor chain we were modeling.[134] Based on prior experimental experience on IL-23 interactions, Dr. Jamie Spangler advised me that the interaction was likely between the D1 domain of IL-23R and IL-23 with the conserved Trp-156 on IL-23B serving as the 'lightning rod'. Using this information, I obtained a starting state and locally docked the receptor against the cytokine heterodimer while constraining the conserved tryptophan residue

to contact the receptor. This was the first target for which I used RosettaDock 4.0, and as a result I was able to efficiently dock 65 receptor formations to 56 cytokine conformations.

Figure 3.5 shows our best model superimposed on the crystal structure (5MZV, in grey). The conserved tryptophan of IL-23B is highlighted in red. This model was able to capture the rough binding mode along with the tryptophan lightning rod interaction. With 22% of the native contacts recovered and an Irmsd of 0.7 Å, our model was adjudged acceptable. None of my models of IL-23R (yellow) had $RMSD_{C\alpha}$ under 4.2 Å because of the different orientations of the three domains and as a result the Lrmsd of the model was 16.8 Å. The best

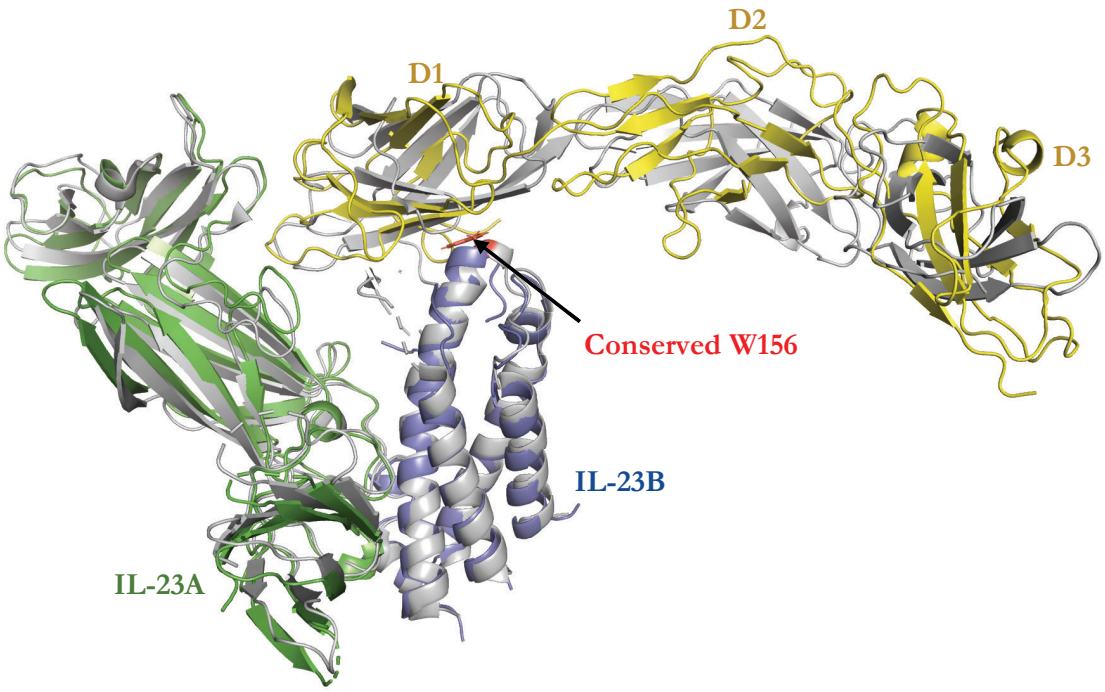


Figure 3.5: T122: our best model of the complex of the two chains of IL-23, viz. IL-23A (green) and IL-23B (blue) with IL-23R (yellow) superimposed on the crystal structure (grey). The lightning rod interaction via the conserved Trp-156 (red) to IL-23R domain 1 (D1) was correctly predicted. IL-23R model had large errors in the relative orientation of the three domains.

model across all the CAPRI groups was a medium-quality model with 40% of the native contacts.

3.3.2. Failures

3.3.2.1. Target 113: Contact-dependent toxin–immunity protein complex

(with Jeliazko R. Jeliazkov)

Gram-negative bacteria use contact-dependent growth inhibition (CDI) systems to deliver toxins arresting growth of competing bacteria when they come in direct contact with them. This system consists of a variable toxin, an outer membrane protein exporting the toxin, and an immunity protein to prevent self-toxicity by recognizing and neutralizing the toxin.[135] In T113, we were asked to model the interaction between the C-terminal domain of the toxin, CdiA-CT, and its cognate immunity protein, CdiI2 from *Cupriavidus taiwanensis*.

We started with monomer models from CASP12 predictors. We observed variability in CdiA-CT models and chose nine of them that had convergent secondary structure signatures. There was less variability in CdiI2 models, but the eleven-residue N-terminal tail had some variety. We chose three models with significantly different tail conformations from each other to hedge our bets. As we could not find a homologous complex, we searched the global conformational space using ClusPro[136] and chose the binding mode compatible with most of the monomer conformations. Restricting our search to the local space around this mode, we

then docked the ensemble of nine CdiA-CT backbone with three CdiI2 backbones to generate 15,000 models.

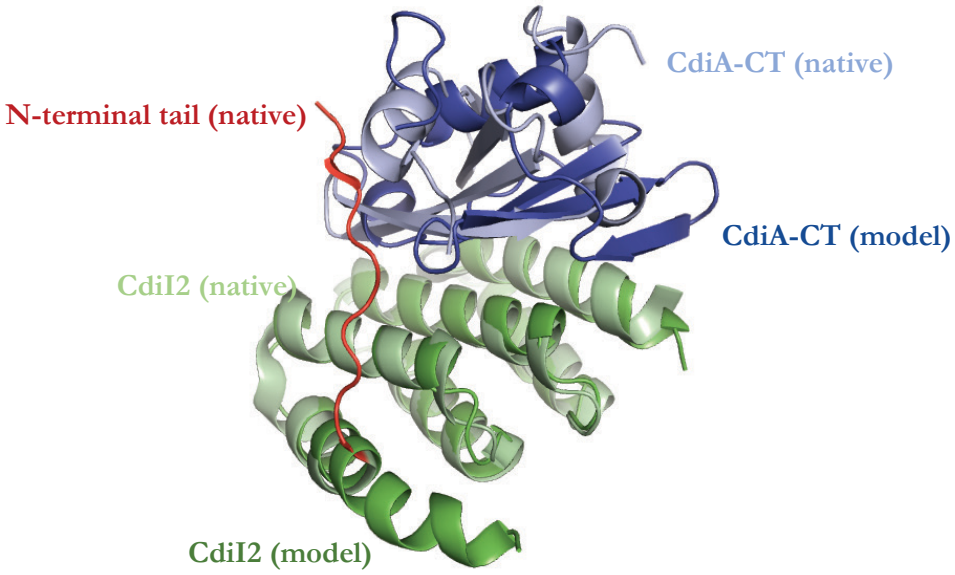


Figure 3.6: T113: Model of CdiA-CT (blue) and CdiI2 (green) superimposed on the crystal structure of CdiI2 (pale green). The N-terminal region of CdiI2 was incorrectly predicted to be helical. Contacts with the N-terminal (red) are required for the proper rigid body orientation of CdiA-CT (pale blue).

Figure 3.6 shows our best model superimposed on CdiI2 in the crystal structure (pale green) of the complex (5T87). The N-terminal tail, which none of the models predicted correctly, is highlighted in red to show contacts it makes with CdiA-CT (pale blue). As we could not predict this tail conformation correctly in the model (green), we predict the rigid body conformation of CdiA-CT incorrectly (blue). The best model across all groups was classified as acceptable.

3.3.2.2. Target 114: Ljungan virus protein (with Naireeta Biswas)

T114 was the homodimeric-protein 2A2 from Ljungan virus. The function of this protein is unknown. We were provided with monomer models from CASP12 participants. We relaxed the models and chose five top-scoring distinct models for docking. We found no homologs from which to extract symmetry information, and hence, we performed a global search of the relevant search space to generate 50,000 models for each monomer. None of the models submitted by any predictor was adjudged to be correct. As the experimental structure of this protein has not been released yet, we could not determine the reason(s) for failure.

3.3.2.3. Target 116: Bifunctional histidine kinase

One of the principal regulators of bacterial cell cycle division is the activity of cell cycle kinases, which are the functional counterparts of cyclin-dependent kinases in eukaryotes. Recently, it was shown that cyclic-d-GMP binds the histidine kinase, CckA, at different stages of the cell cycle and switches it from a kinase to a phosphatase (or vice versa) driving cell cycle progression.[137] T116 was two domains (Dhp and CA) of the homo-dimer, CckA of *Caulobacter crescentus*. I identified several homo-dimeric histidine kinase homologs with 97% or more sequence coverage and 25% of more identity like those from *E. coli* and (4GCZ) and *Geobacillus stearothermophilus* (3D36). Each of them had a different relative orientation of the Dhp and CA domain counterparts. Therefore, this target could only be successfully docked if the domains were correctly oriented in the monomers.

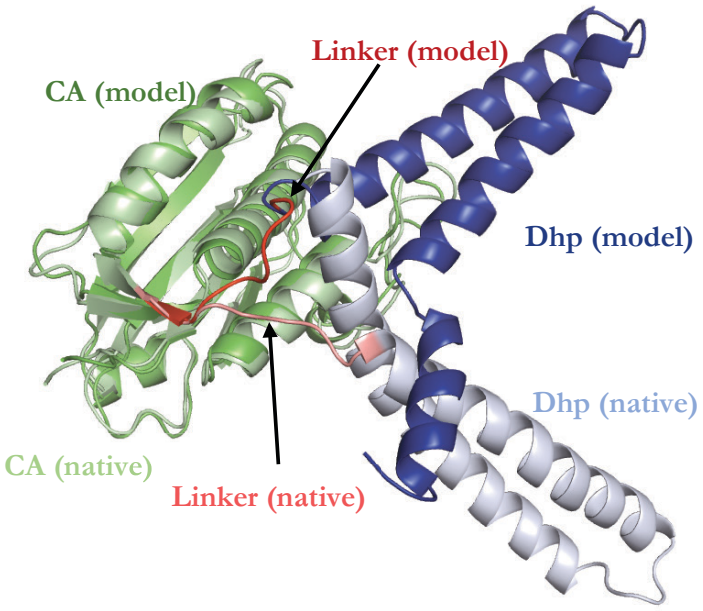


Figure 3.7: T116: Monomer model CckA with the CA domain (pale green) superimposedto that of the crystal structure (pale green). The incorrect conformation of the linker region in the model (red) displaces the Dhp domain (blue). The correct linker conformation (salmon) is required for the correct placement of Dhp (pale blue), and eventually the dimer.

Models from CASP12 participants had a variety of different relative orientations of these two domains depending on the homolog template they chose. Using monomers of two different orientations, I generated 25,000 docked models per orientation.

Unfortunately, the relative orientations of the two CckA domains were very different from all available homologs. Figure 3.7 shows an overlay the crystal structure and one model of the CckA monomer. Upon superimposing the model CA domain (green) on the native one (dull green), we see that the linker connecting the domains in the model (red) and the native (salmon) have completely different conformations. This changes the position of the Dhp

domain of the model (blue) compared to the native (dull blue). Without a good monomer conformation, we, as well as all the other predictors, failed to dock the dimer correctly.

3.3.2.4. Target 117: Pins–Insc tetramer

Mammary stem cells in adult humans drive the reshaping and regeneration of mammary glands during puberty and pregnancy. To position their daughter cells to differentiate into specialized cells types, they undergo asymmetric cell division. The localization of a protein called Pins determines cell polarity by forming a complex that tethers the mitotic spindle. The adaptor protein Insc binds Pins and prevents the formation of the tethering complex.[138] Instead, the Pins–Insc complex recruits cell fate determining proteins. Thus, asymmetric distribution of Insc causes differential organization of cell fate determinants in the daughter cells. T117 was the tetrameric complex of two Pins and two Insc units.

A structure of the Pins monomer was available (3SF4). For the structure of Insc, I relied on models from CASP12 participants. Owing to the absence of close homologs, I obtained a variety of different models, which I then relaxed and clustered by similarity. The models that clustered most tightly still had a variety of conformations of 35 residues at the N-terminus, which I consequently truncated. Based on literature,[138] I decided to take a homo-dimer of hetero-dimer approach, where I first docked the Insc to Pins and generated 50,000 models. Selecting an ensemble of 15 distinct top-scoring dimers, I symmetrically docked the dimers starting from four distinct orientations and selected from 50,000 models of each orientation.

Figure 3.8 shows the crystal structure of the complex (5A7D) with two units of Pins (in shades of blue) and two units of Insc (in shades of green). This structure is a homo-dimer of two hetero-dimers as I predicted, but is not symmetric. The primary contacts of Insc in each hetero-dimer unit occurs in the thirty-residue unfolded N-terminal peptide, $Insc^{PEPT}$ (highlighted in red). As a result, there is a large amount of conformational flexibility in the hetero-dimer subcomplex with the two dimers in the crystal structure having significantly different conformations. I had truncated this peptide and hence could not model either the hetero-subcomplex or the whole complex correctly. This was arguably the hardest challenge of round 37 because it involved not only multi-body docking, but also predicting the interactions of an unfolded peptide stretch with large conformational flexibility. No CAPRI team submitted an acceptable or better model.

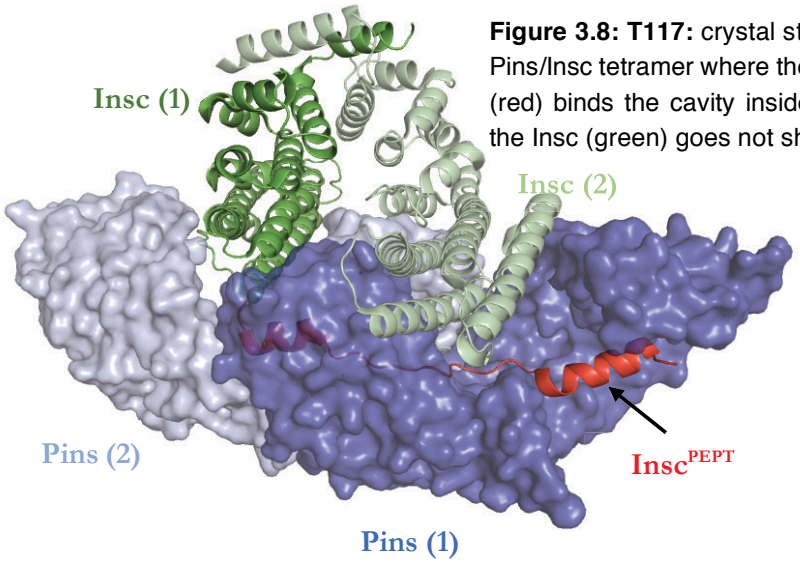


Figure 3.8: T117: crystal structure of the asymmetric Pins/Insc tetramer where the N-terminal region of Insc (red) binds the cavity inside Pins (blue). The rest of the Insc (green) goes not show a strong interaction.

3.3.2.5. Targets 123 & 124: PorM–camelid nanobody complex (led by Jeliazko R. Jeliazkov)

Type IX secretion systems are used by bacteria to secrete cell-surface adhesins that facilitate motility on solid surfaces. In *Porphyromonas gingivalis*, one of the periplasmic members of this complex is PorM.[139] Recently, fragments of single chain antibodies from camelids called nanobodies have drawn interest as chaperones for crystallizing proteins that are difficult to crystallize otherwise.[140] T123 and T124 are the N- and C- terminal domains of PorM in complex with nanobody chaperones, respectively. While the N-terminal domain was crystallized as a monomer, the C-terminal domain was crystallized as a dimer.

Nanobodies (nb) recognize antigens by interacting with them through three variable loops called H1, H2 and H3. The H3 loop is the longest and most flexible loop, and as a result it is the primary determinant of complementarity. In T123, the H3 loop of nb02, which was complexed with the $PorM_{N\text{-}term}$, was 12 residues long, and in T124, the H3 loop of nb130, which was complexed with $PorM_{C\text{-}term}$ dimer, was 21 residues long, which makes it difficult to thoroughly sample their conformations. First, we modeled the constant core of the nanobody and the H1 and H2 loops from available homologs. Next, we generated 1,000 models with different H3 loop conformations using RosettaAntibody.[141,142]

For T123, we obtained $PorM_{N\text{-}term}$ models from Robetta. For T124, we obtained $PorM_{C\text{-}term}$ monomer conformations from Robetta and docked them together symmetrically to attain a dimer configuration. In both cases, no homologs were available as templates and hence the

monomers were modeled *de novo* from the sequence, which introduces larger errors. Using the best-scoring PorM models, we searhed for suitable nanobody-binding regions by global docking using ClusPro. We then refined the distinct binding modes obtained from ClusPro while simultaneously sampling various nanobody variable loop conformations using SnugDock,[124,141] a variant of RosettaDock specialized for docking antibodies.

The structure of T123, i.e. the $PorM_{N\text{-}term}$–nb02 complex is not yet released and we cannot analyze the reason for our failure. Only one acceptable solution was submitted across all participants. T124, i.e. the $PorM_{C\text{-}term}$ dimer–nb130 complex involved multiple challenges: modeling monomers without templates, dimerizing them and then docking the nanobody correctly. Figure 3.9 shows the crystal structure of this complex (6EY0) with the two $PorM_{C\text{-}term}$ subunits in shades of green and nb130 in blue. The H3 loop of nb130 from a cross-beta sheet with one of the $PorM_{C\text{-}term}$ subunits, which is highlighted in red. In this crystal structure, the other $PorM_{C\text{-}term}$ subunit also has a nb130 monomer bound, thus making the assembly tetrameric. All our $PorM_{C\text{-}term}$ monomeric models were incorrect, and as a result, although we identified an nb02 H3 loop with 2.4 Å $RMSD_{C\alpha}$, all our docking models had an Lrmsd of more than 18 Å. As a result, we, as well as all the other participants, failed to model this complex correctly.

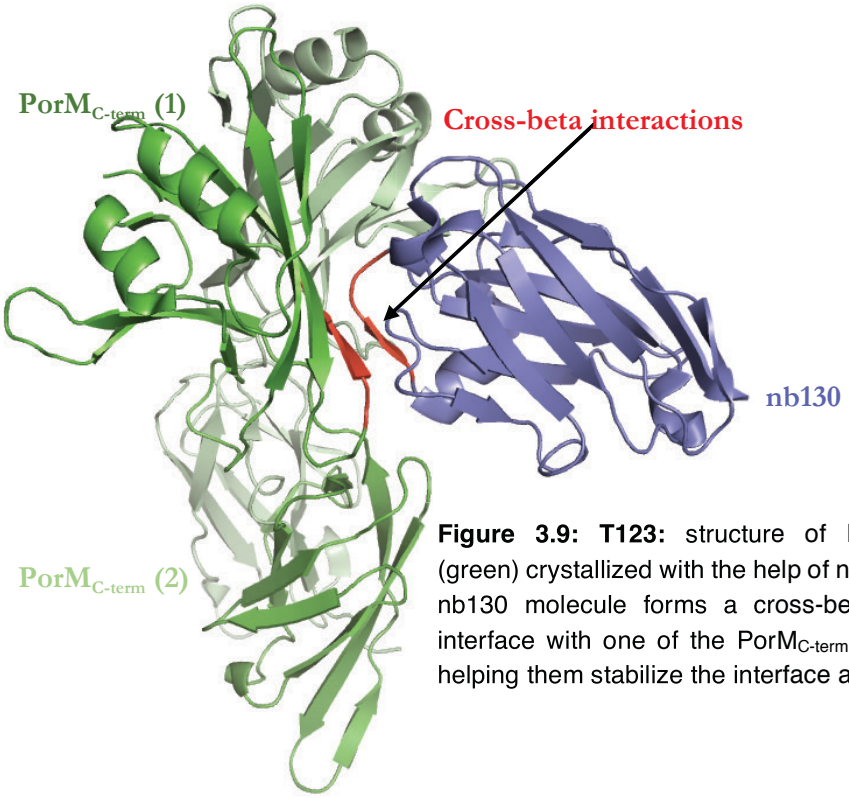


Figure 3.9: T123: structure of PorM$_{C\text{-}term}$ dimer (green) crystallized with the help of nb130 (blue). One nb130 molecule forms a cross-beta sheet at the interface with one of the PorM$_{C\text{-}term}$ molecules, thus helping them stabilize the interface and crystallize.

3.3.2.6. Targets 131 & 132: CEACAM1–HopQ complex

(led by Morgan L. Nance)

A leading cause of gastritis and gastric ulcers is *Helicobacter pylori* infection. These bacteria target cell adhesion molecules on the gastric epithelium, CEACAMs, by attaching onto them using bacterial adhesins, HopQs.[143] T131 and T132 were the complexes of HopQ1 and HopQ2, respectively bound to the N-terminal domain of CEACAM1.

Multiple structures were available for the N-terminal domain of CEACAM1 (2GK2, 4QXW, 4WHD, and 5DZL).[144,145] For T131, the structure of HopQ1 was available with four

loops missing at the putative binding interface (5LP2).[143] Using this structure as a template, complete models were obtained from Robetta. Robetta produced different conformations for the two longest loops (residues 135–148 and 245–255), which suggested potential for flexibility. A mutation study indicated that residues Y34 and I91 of CEACAM1 are essential for HopQ binding.[146] The authors of the study also conjectured that the first long loop of HopQ1 is involved in binding CEACAMs. We modeled the missing loop using fragment insertion and closed the loop with cyclic coordinate descent.[147] Using an ensemble of 200 different loop conformations for the first HopQ1 loop and constraints to ensure CEACAM1 Y34 and I91 contact HopQ1, we generated 10,000 models each from two different starting states. For T132, we modeled the structure of the HopQ2 monomer based on its homology to HopQ1 using Rosetta Remodel.[148] We followed a similar protocol for HopQ1 loop conformation sampling (for a slightly shorter loop of residues 135–144) and docking.

In both the cases, our loop modeling methods failed to provide the necessary bound conformation, often producing extended loops, instead of the compact structure in the crystal. Figure 3.10 shows the crystal structure (6GBG) of HopQ1 (green) bound to the N-terminal domain of CEACAM1 (blue). The first loop of HopQ1 (highlighted in red) adopts a strand-turn-helix motif that we did not sample. As a result, the rigid-body orientation of CEACAM1 was completely incorrect. In retrospect, it would have been better to take an approach alternating loop modeling and docking, as is done in the antibody docking protocol SnugDock. The two CEACAM1 residues predicted to be at the interface were indeed found to be there, and are shown as salmon sticks.

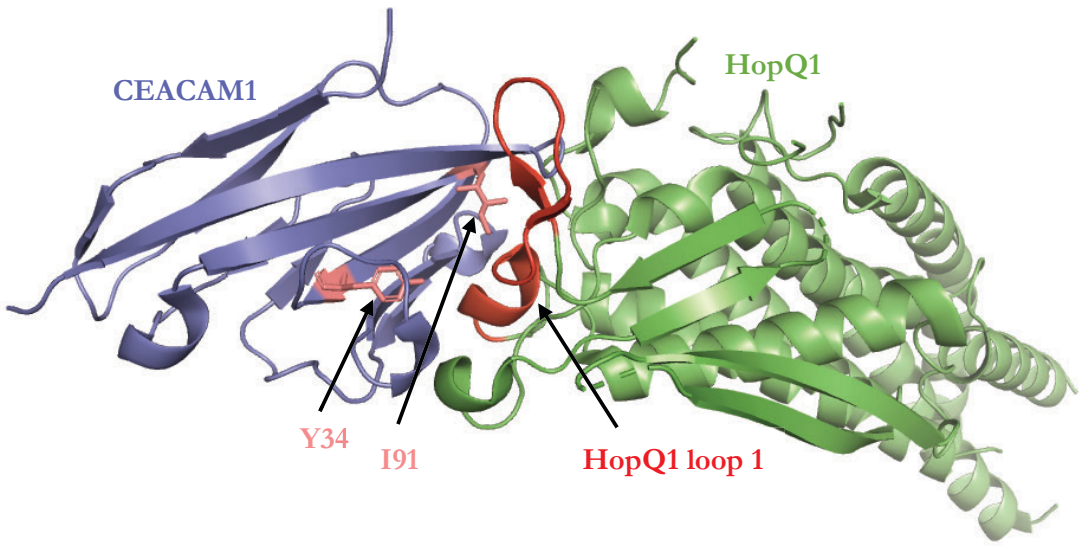


Figure 3.10: T131: crystal structure of HopQ1 (green) bound to CEACAM1 (blue). The 14-residue loop of HopQ1 (red) is stabilized by interactions with CEACAM1, including the two residues predicted to be at the interface (salmon). The strand-turn-helix conformation adopted by the loop was not predicted correctly leading to incorrect rigid-body placement of CEACAM1.

While we did not predict the structure correctly, we did successfully refine and score structures submitted by another group. Our best refined model was classified as acceptable with 27% of native contacts predicted, Lrmsd and Irmsd of 11.8 Å and 3.2 Å, respectively. This demonstrates that the *REF2015* score function can recognize the near-native structure. Therefore, the challenge is to sample the conformation *de novo*.

3.3.3. To be announced

For the following targets, CAPRI results and experimental coordinates have not yet been released at the time of writing, so it is not possible to comment on the accuracy of the models.

3.3.3.1. Target 115: Receptor-binding domain of virus (with Joseph H. Lubin)

The whitewater arroyo virus infects a host cell by attaching to the transferrin receptor and gaining entry.[149] T115 was the receptor binding domain, GP1, of the glycoprotein complex on the viral surface. The oligomeric state was given to be homo-dimeric, which we presumed to mean symmetric homo-dimeric. The monomer models from CASP12 participants were convergent, which gave us confidence in using them. As no dimeric homolog was found, we globally sampled the conformational space and generated 50,000 docked models.

Although the results for this target were not announced, the crystal structure was available (5NSJ).[149] Surprisingly, the interaction between the two proteins was asymmetric. While the crystallographers concluded that the asymmetric unit consists of the asymmetric dimer, they did not claim the dimer to be a biological assembly.[150] Moreover, homologs of this protein bind to the transferrin receptor as monomers. Based on the further analysis using EPPIC,[151] I speculate that the asymmetric dimer was an artifact of crystallization rather than a biological assembly; crystal contacts were accidentally identified as biological association for this target.

3.3.3.2. Target 125: NKR-P1–LLT1 hetero-hexamer

Natural killer cells provide innate immunity by recognizing pathogens through a variety of cell surface receptors including NKR-P1. LLT1, a cell surface ligand of NKR-P1, is presented by other cells in the body for self-recognition.[152] T125 was the complex between the extracellular domains of NKR-P1 and LLT1. This was a three-step docking challenge: first, a

dimer of NKR-P1 had to be modeled, then the LLT1–NKR-P1 dimer complex had to be determined, and finally, two of these hetero-trimers had to be docked together to construct the hetero-hexamer.

For NKR-P1, I generated dimer models by symmetric docking from monomer models of NKR-P1 obtained from Robetta. I chose seven dimer configurations for further docking. On *post ex facto* analysis, the closest docked conformation had an $RMSD_{C\alpha}$ of 4.7 Å from the crystal structure of NKR-P1 dimer (5MGS). I then modeled the NKR-P1 dimer–LLT1 complex by global docking models using ClusPro and locally refining them in Rosetta. A structure of LLT1 dimer was already available (4QKH); I used this as a reference to place the whole complex together. This step also acted as a filter to weed out trimer configurations that clashed with each other. Finally, we locally refined three candidate complexes to generate 5,000 models each.

3.3.3.3. Targets 126–130: Arabino-oligosaccharide binding to proteins

(led by Dr. Jason W. Labonte and Morgan L. Nance)

A pivotal step in the carbon cycle is the degradation of plant biomass by soil bacteria.[153] These bacteria break down the plant cell wall by digesting complex polysaccharides. One such polysaccharide, L-arabinan, which is composed of a variety of arabinosaccharide units, is recognized and digested by the L-arabinan-utilization system of bacteria like *Geobacillus stearothermophilus*.[154] In round 41 of CAPRI, we modeled the interaction between two important components of this system—the arabinose sensor, AbnE, and the arabinanase, AbnB—and

arabino-oligosaccharide ligands of different lengths. Specifically, T126–129 challenged us with the docking of 1,5-α-L-arabinohexose (A6) through 1,5-α-L-arabinotriose (A3), respectively, to AbnE. T130 involved the docking of A5 to a catalytic mutant (E201A) of AbnB.

We modeled AbnE from homologs with 95% or more sequence coverage and 25% or more identity using Modeller[132] and relaxed the models in Rosetta.[107] Alternatively, we also obtained models from the Robetta server.[133] One of the homologs that we used to model the target, the maltose-binding protein GacH from *Streptomyces glaucescens*, exists in two conformations: an unliganded open conformation and a closed, ligand-bound conformation.[155] From all the aforementioned protein models, we used the conformation closest to the ligand-bound GacH conformation to model T126–129. As the chemical description of arabinose was absent in Rosetta, we programmed the required geometry, partial charge, and chemical connectivity information to model arabinose ligands. To obtain a starting structure, we superimposed the AbnE model and A4 onto maltotetraose-bound GacH (3K00) by changing the backbone torsion angles of A4 to best align with maltotetraose. For A5 and A6, we added arabinose units to the non-reducing end of the ligand. For A3, we removed an arabinose unit from the non-reducing end.

To dock the glyco-ligands while exploring their various backbone conformations, we used the new GlycanDock protocol in Rosetta.[156] In this protocol, the glyco-ligand undergoes small backbone motions along with rigid-body moves to fit the protein cavity. These perturbations are alternated with side-chain repacking of the protein residues at the protein–glycan interface. For each target, we obtained 15,000 initial docked models without any constraints to relieve

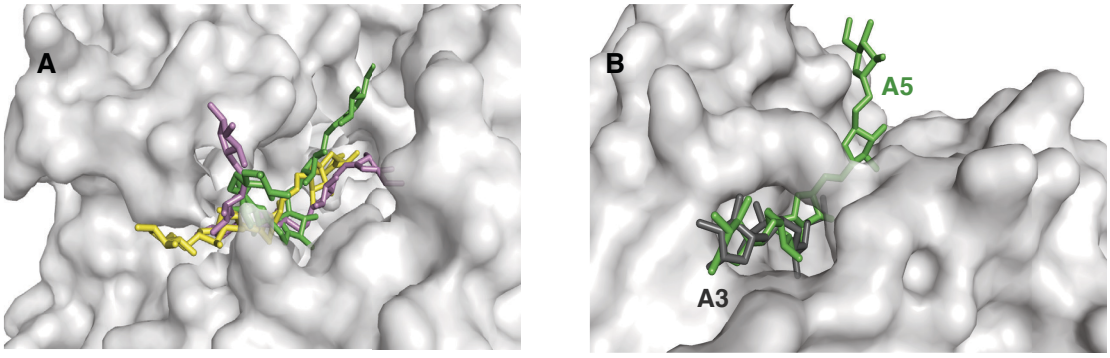


Figure 3.11: (A) **T126:** range of ligand conformations sampled by 1,5-α-L-arabinohexose (green/yellow/pink) in the binding groove of AbnE. (B) **T130:** a predicted conformation of 1,5-α-L-arabinopentose (green) on $AbnB_{E201A}$. The crystal structure of 1,5-α-L-arabinotriose (dark grey) is shown for comparison.

clashes and to broadly sample the rigid-body conformational space. From the models where the ligands had less than 5 Å RMSD from the starting structure, *i.e.* those that stayed in the binding pocket of AbnE, we selected the one with the lowest interaction energy as the starting model for the final simulation. For the final docking simulation, we added constraints to hold the glyco-ligands within the putative binding pocket of AbnE and generated another 15,000 models. The range of conformations explored by A6 in T126 is shown in Figure 3.11A.

For T130, the crystal structure of A3 bound to the E201A mutant of the glycosidase AbnB was already available (3D5Z). The active site of this enzyme is a long groove with a bridge connecting the brinks under which the ligand can slide (see Figure 3.11B). The groove has no steric obstruction at either end to hold the substrate in place and cleave a particular glycosidic linkage, and hence it cleaves linkages indiscriminately.[154] Consequently, although a structure was available with A3, we could not *a priori* predict how the A5 ligand would position itself.

Extending the A3 in either direction provided us with starting coordinates. We generated 10,000 docked models each from four starting states using GlycanDock while constraining the A5 ligand to the active site groove. Figure 3.11B shows one of the predicted conformations of A5 (in green) superimposed on the crystal structure of AbnB (light grey)–A3 (dark grey) complex.

3.3.3.4. Target 133: Colicin DNase–immunity protein complex (led by Morgan L. Nance)

Similar to the proteins modeled in T113, T133 was another toxin–immunity protein pair. In this case, it was the complex of the DNase, colicin, that *E. coli* releases due to environmental stress and its cognate immunity protein, Im. What makes this an interesting model system is that Im binds Colicin over a millimolar to femtomolar range where a single residue change can greatly alter affinity.[157] T133 was a colicin E2 DNase–Im2 complex designed to change partner specificity from the native complex.

The crystal structure of the native colicin E2 DNase–Im2 was available (3U43).[157] However, the organizers informed us that the mutations lead to an altered binding mode. Therefore, the challenge of this target was recognizing changes in binding mode brought about mutations. The designed colicin, EDes3 had mutations in 17 of the 132 positions while the immunity protein ImDes3 had 15 of its 85 positions mutated, most of which were situated on a loop. Three residues identified as native-sequence hotspots for binding (Y54 and Y55 on Im2 plus F86 on E2)[157] were not changed. After mutating and refining the structure of the

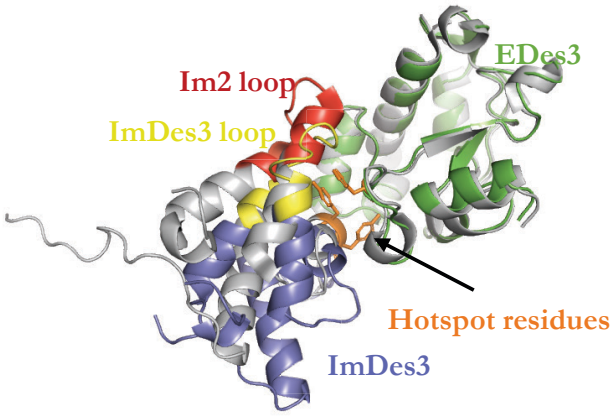


Figure 3.12: T133: a model of ImDes3 (blue) bound to EDes3 (green) superimposed on the crystal structure of the wild type E2/Im2 complex (gray). The Im2 loop (red) is mutated to become the ImDes3 loop (yellow). This alters the rigid body orientation of ImDes3 by pushing it 'down'. The hotspot resides across the interface in the wild type (orange), *viz.* Y54-Y55 on Im2 and F86 on E2 still interact in the model of the designed complex.

mutant proteins, we explored different conformations of the ImDes3 loop with the mutations (residues 20-35) and closed the loop with kinematic closure.[158] For EDes3, we obtained a variety of backbone conformations using Rosetta Backrub.[108] We then docked an ensemble of EDes3 conformations with an ensemble of ImDes3 conformations while constraining the three hotspot residues to interact.

Figure 3.12 shows our top-scoring Edes3 (green)–ImDes3 (blue) model superimposed with the crystal structure (gray) of E2–Im2 complex (3U43). We predict a different binding mode due to changes in the loop between Im2 (in red) and ImDes3 (yellow). The hotspot residues (orange sticks) still interact despite a change in the overall binding mode.

3.3.3.5. Target 136: Lysine decarboxylase homo-decamer

Inducible decarboxylases in enterobacteria like *E. coli* help counteract acid stress by producing polyamines from lysine, arginine and ornithine.[159] T136 was the homo-decameric lysine decarboxylase, LdcA from *Pseudomonas aeruginosa*. Close homologs were available for the

complex in the form of lysine decarboxylase, LdcI from *E. coli* (5FKZ) and arginine decarboxylase, AdiA (2VYC) from *Salmonella typhimurium*, both of which had 85% or more sequence coverage and 40% or more identity as well as the same point symmetry. The organizers informed us that the wing domain of the subunits was significantly different to the homologs leading to different inter-subunit contacts. As the subunit arrangement was likely to be similar to its homologs, the challenge of this complex was to model the wing domain correctly within the confines of this D5 symmetry. Another issue was the sheer size of the protein: ten subunits each with 750 residues making extensive interfaces with other subunits.

I started by modeling the monomer using Robetta, which produced convergent conformations for the wing domain that were distinct from the homologs. Drawing symmetry information from the homologs and arranging the subunits accordingly led to steric clashes at the wing domain. All fixed-backbone refinement efforts using SymDock failed to produce a structure free of clashes. SymDock refinement ended up expanding the complex by increasing the inter-subunit distance to relieve the clashes. (I had observed this tendency in T110 and T118, but they did not have as extensive an interface as this complex and hence, the effect of this expansion was less consequential in those cases.) To keep the complex together, we tried reducing the weight of the repulsive term of the van der Waals term of Rosetta's score function. While this kept the complex together, it caused major steric clashes.

I tried producing alternate monomer conformations by generating ensembles by relaxing the monomer. However, for each clash relieved, a new clash was found. I conjectured that the monomers needed to be relaxed in the context of the complex, and not independent of it. On

doing so and then docking with the context-refined monomers, I was able to obtain structures where the wing domain readily fits into the given symmetry without steric clashes. **This target was the basis of the flexible-backbone refinement protocol for symmetric proteins described in Chapter 4.** Figure 3.13 shows one of our models for LdcA with the wing domains of five subunits highlighted in darker hues against lighter tints of the rest of the subunit. I predict that each wing domain contacts two neighboring wing domains and well as a neighboring chain.

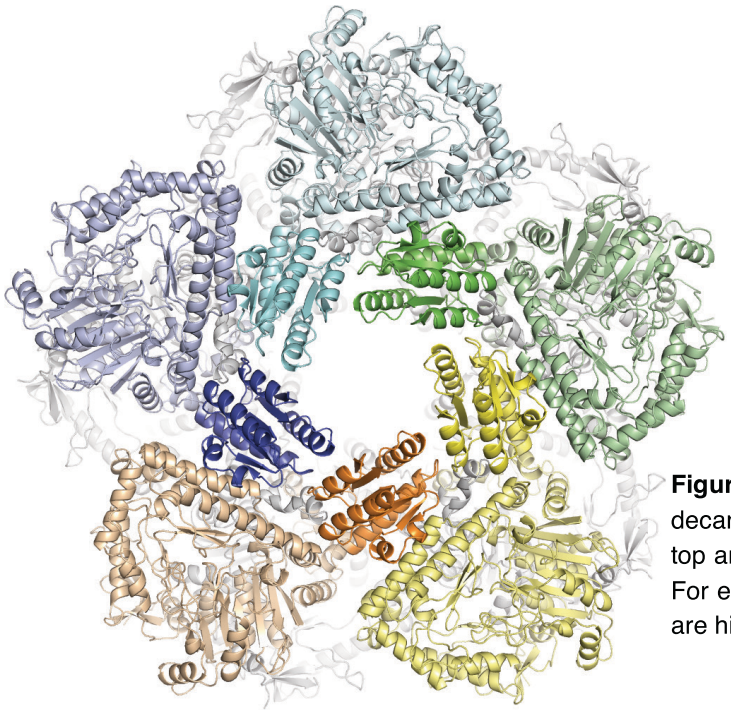


Figure 3.13: T136: a model of LdcA decamer where the five subunits on top are displayed in different colors. For each subunit, the wing domains are highlighted in brighter hues.

3.4. Discussion

Previous rounds of CAPRI led to the development of niche protocols like SnugDock[124] to model antibody-antigen binding and pHDock[128] to dynamically sample residue protonation states while docking.[121] In rounds 37–45, we utilized these specialized methods while also encountering challenges that require overhauls of the core methodology for general problems such as global docking with flexibility, global docking of symmetric homomers and oligosaccharide-protein docking. We modelled backbone flexibility by incorporating pre-generated ensemble of backbone conformations during docking. With RosettaDock 4.0, I sampled over fifty conformations for each partner to successfully model T122. Despite having an efficient backbone sampling algorithm, we failed to model T131 and T132 due to the absence of conformations where the interacting loops were in near-bound conformation. These failures highlight the need to develop ensemble generation methods that sample loop conformations broadly.

As many of the targets were symmetric homomers with varying degrees of homology to existing structures, we were able to thoroughly assess Rosetta's SymDock protocol. When homologs were present, I could borrow the symmetric arrangement from the homolog as a template, as I did to successfully model targets 110, 111, 112, 118 and 119. However, even in those cases, the proximity of the monomer backbone to the template monomer backbone determined the overall quality of the models. For example, the monomer model of T111 had a 0.8 Å $RMSD_{C\alpha}$ from the template and was our only prediction to be classified as high-quality.

While one would expect that the more closely related a template is, the better the model will be, I noticed a systematic pattern of error.

When starting from a symmetric arrangement based on a distantly-related template, the complex would become less compact. How much the complex expanded was a function of how tightly packed the complex was: for the trimeric T110, the inter-subunit distance increased by about 1 Å, for the octameric T118, it increased by over 4 Å, and for the decameric T136, it increased so much that all starting residue-pair contacts were lost. I conjectured that imposing symmetric arrangements drawn from homologs on monomers modeled independently led to clashes that SymDock's fixed-backbone refinement was unable to resolve. To circumvent this issue for T136, I relaxed one subunit after positioning it in the putative complex. When I reassembled the complex with this monomer conformation—even though a major interacting domain (wing domain in Figure 3.8B) was different from the template—SymDock was able to find a decamer model whose inter-subunit distance was 62.2 Å compared to 61.6 Å in the template. This approach inspired the flexible-backbone refinement strategy of the new symmetric docking protocol described in the next chapter.

For only the second time in CAPRI, we encountered oligosaccharide–protein complexes. Five targets in round 41 gave us an opportunity to work with the recently-developed RosettaCarbohydrate framework[156], especially the GlycanDock application therein. With an additional mobile backbone torsion, multiple mobile side chains, and flexible rings, oligosaccharides have many more degrees of freedom than peptides do. GlycanDock samples these mobile dihedrals while performing rigid-body transformations to place the

oligosaccharide in a binding pocket while simultaneously repacking the side chains of contacting protein residues. Despite extensive sampling, we recognized a deficiency in GlycanDock, *viz.* its inability to hold the glyco-ligand in the binding pocket. We realized that large conformational moves were causing the ligand to clash with the protein and hence, reduced the magnitudes of the moves. Thus, we used constraints to favor conformations that stayed in the binding pocket. This experiment revealed the need to optimize the GlycanDock protocol with a variety of glyco-ligands, which my colleagues in the Gray laboratory are currently doing.

With seven successful predictions and one additional scoring success, our performance in the rounds evaluated thus far was on par with other leading groups. Of our seven docking failures, I believe, retrospectively, that we had the sampling techniques available in Rosetta to better model targets 113, 131, and 132. On the other hand, targets 114, 116, 117, and 124 required blind prediction tools that do not yet exist and as a result, they did not elicit a successful model from any predictor. Broadly, the challenges that caused the most failures were docking with large conformational changes and multi-body docking (especially higher order heteromers). These community-wide failures highlight the massive gaps that still need to be plugged to fulfill the overarching goal of reliably modeling entire interactomes.[160–162]

Chapter 4

Flexible backbone assembly and refinement of symmetrical homomeric complexes

[Pre-print version available as Roy Burman, S. S. , Yovanno R. A., & Gray, J. J. Flexible backbone assembly and refinement of symmetrical homomeric complexes. *bioRχiv* (2018); doi:10.1101/409730. This chapter contains minor revisions to the pre-print version.]

4.1. Overview

Symmetrical homomeric proteins are ubiquitous in every domain of life, and information about their structure is essential to decipher function. The size of these complexes often makes them intractable to high-resolution structure determination experiments. Computational docking algorithms offer a promising alternative for modeling large complexes with arbitrary symmetry. Accuracy of existing algorithms, however, is limited by backbone inaccuracies when

using homology-modeled monomers. Here, I present Rosetta SymDock2 with a broad search of symmetrical conformational space using a six-dimensional coarse-grained score function followed by an all-atom flexible-backbone refinement, which I demonstrate to be essential for physically-realistic modeling of tightly packed complexes. In global docking of a benchmark set of complexes of different point symmetries—staring from homology-modeled monomers—I successfully dock (defined as predicting three near-native structures in the five top-scoring models) 19 out of 31 cyclic complexes and 5 out of 12 dihedral complexes.

4.2. Introduction

The pervasive appearance of symmetrical homomeric proteins across all domains of life has been attributed to increased stability,[163] fine-tuned functional regulation,[164,165] better synthesis error control,[64] reduced aggregation,[166] and genome compactness.[167] While these evolutionary forces drive proteins towards larger assemblies, the size of these complexes makes it particularly difficult to obtain high-resolution structures. Despite an estimated 50–70% of proteins being symmetric homomers,[51] less than 40% of proteins in the Protein Data Bank[117] are symmetric (as of June 2018). This gap could be bridged by the development of computational docking methods to obtain accurate models of symmetric homomeric complexes.

Especially desirable are versatile methods that incorporate different kinds of experimental data in the modeling pipeline. For low-resolution cryo-EM, NMR or SAXS data, symmetry-constrained flexible refinement is essential for obtaining high-quality models.[168,169] In the

absence of such a model, the method should be able to dock a homology modeled monomer.[133,170,171] This monomer model can be combined with experimental determination of the oligomeric state and/or symmetrical placement of subunits in homologs to prepare a preliminary model for refinement. If the relative orientations of the subunits cannot be obtained from homologous structures, the method should be able find the correct arrangement of the subunits while restricting the search space to relevant symmetrical conformations.

The symmetry framework in the Rosetta Macromolecular Modeling Suite allows modeling of complexes with arbitrary symmetries.[172] The framework has been used to develop SymDock, a docking protocol for point symmetries. SymDock has been shown to correctly model complex structures from a monomer for a variety of symmetry groups.[67] This protocol uses a coarse-grained phase to sample symmetric conformation space starting from a random or a pre-defined orientation followed by an all-atom phase for refinement. To further improve models, the Rosetta suite also allows integration of information from a plethora of experimental methods like cross-linking studies,[173] NMR,[174] and SAXS[175] as well as co-evolutionary analysis[176] while docking. This two-phase approach and the variety of ways of adding constraints make SymDock an extremely versatile tool.

In the last published rounds of the blind docking challenge, Critical Assessment of PRediction of Interactions (CAPRI), although multiple groups generated high-quality models for various homodimers, no group was able to predict high-quality models for the five homotetramer targets, including two for which no acceptable solutions were submitted.[72]

Recently, four leading symmetric docking methods were evaluated on a benchmark of 251 complexes, 180 of which were homodimers.[69] Despite a favorable benchmark composition, starting with homology-modeled monomers, none of the methods was able to produce a CAPRI-acceptable model in the top ten predictions for more than half the complexes. For Rosetta SymDock, I have been aware of two limitations preventing consistently accurate predictions. First, the scoring method used in the coarse-grained phase does not sufficiently discriminate between near-native and spurious interfaces. Second, in tightly packed complexes, small steric clashes between the subunits were not being resolved. Specifically, I observed that symmetric side chain packing and minimization were inadequate for resolving clashes between subunits in tightly packed complexes; such cases additionally require small backbone motions.

In this study, I address these limitations and demonstrate a drastically improved performance of this protocol. First, to enhance model evaluation in the coarse-grained phase, I employ a fast and accurate scoring scheme called Motif Dock Score (MDS). We had previously developed MDS for docking heterodimeric complexes, and it greatly increased the number of conformations with near-native interfaces after a coarse-grained search.[79] Second, I test two approaches to backbone flexibility that have been successfully used for heterodimeric complexes, *viz.* imitating conformational selection[85,86,88] and induced fit.[60,61,115] For conformational selection, I pre-generate an ensemble of conformations from the monomer and used them as input monomers for docking. For induced fit, I minimize energy along the backbone dihedrals and repack side chains during refinement, starting with a low

repulsion between the atoms and progressively ramping it up. The refinement is performed after the rough subunit arrangement has been predicted in the coarse-grained phase.

I evaluate the enhanced protocol, SymDock2, on a diverse benchmark of 43 complexes belonging to the two most common symmetry groups, cyclic (described by a single rotational symmetry axis) and dihedral (described by a rotational symmetry axis and a perpendicular axis of two-fold symmetry). As these proteins rarely crystallize as monomers, I use monomers predicted by a homology docking server as a proxy for the 'unbound' structure. Given a particular point symmetry, I perform a global search of the relevant symmetrical conformation space. These inputs represent the most difficult case described earlier where the monomer conformation is approximate and the subunit arrangement is unknown. This workflow is similar to one commonly employed in CAPRI blind docking.[105] The performance for both the coarse-grained and the all-atom phases show marked improvements over the original SymDock protocol without compromising the overall speed of the protocol.

4.3. Results

Rosetta SymDock is a Monte Carlo-plus-minimization protocol[75] that models symmetric homomeric complexes starting from a monomer structure and a symmetry definition.[67] Symmetry definitions contain information about the rigid-body arrangement of the subunits, how to yield the energy of the whole complex from calculations on one subunit (or a set of subunits), and what the degrees of freedom are along which the subunits are allowed to move.[172] For local docking, specific symmetry definitions can be recapitulated from a PDB file

of a complex whereas for global docking, general symmetry definitions can be loaded for any given point symmetry. In the first, coarse-grained phase of the SymDock protocol, side chains are approximated as 'pseudoatoms'. Coarse-graining allows the subunits to sample the symmetrical rigid-body conformations in a smoothened energy landscape. Next, the side chains are reintroduced and the putative encounter complex is refined by symmetrical side-chain optimization at the interfaces with minimal rigid-body motion. The protocol is illustrated as a flowchart in Figure 4.1.

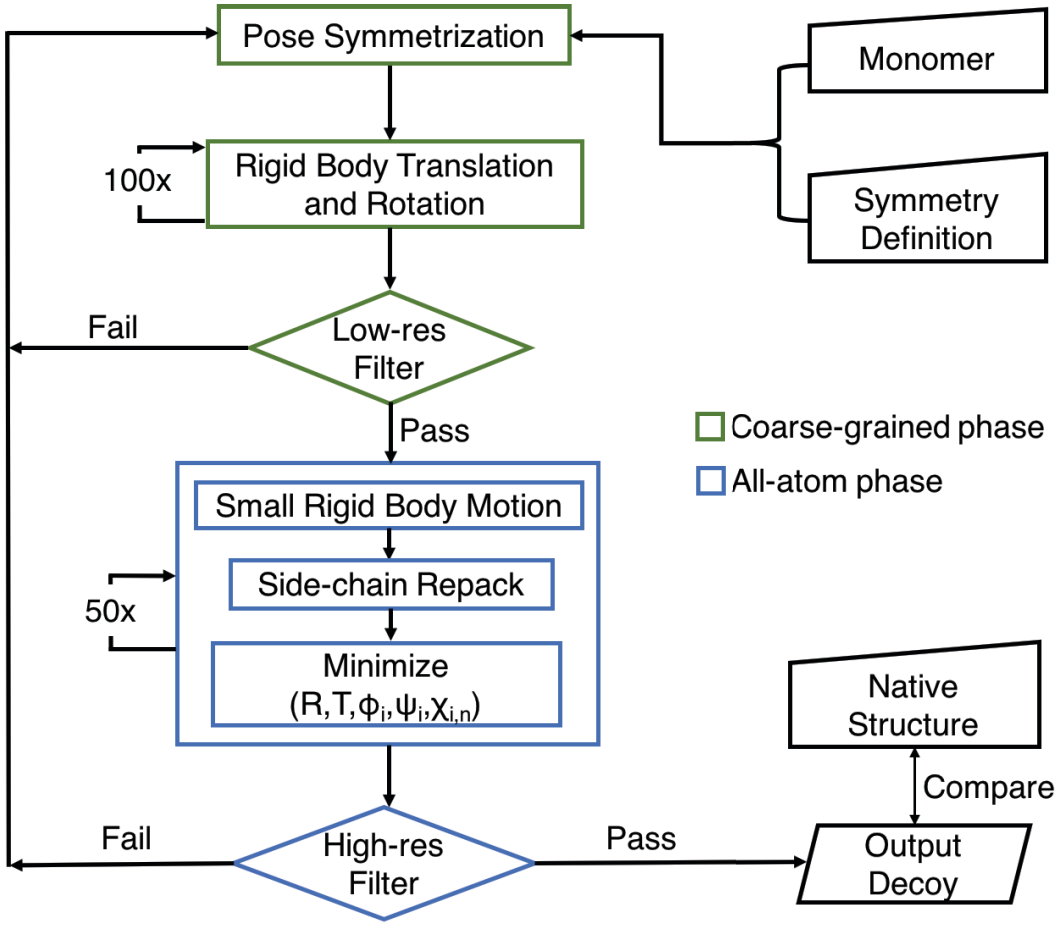


Figure 4.1: Flowchart describing major steps in Rosetta SymDock protocol. In the all-atom phase the structure is minimized along rotational rigid body coordinates (R), translational rigid body coordinates (T), and the dihedrals of the interface residues ($\{\phi_i,\psi_i,\chi_i,n\}$, where $i \in$ interface). See section 4.5.8 for filter descriptions.

4.3.1. Motif Dock Score discriminates near-native interfaces

First, I sought to produce low-scoring, near-native conformations by the broad, coarse-grained search of the symmetrical conformation space. To recognize a near-native conformation, the various interfaces between the subunits must be scored accurately. An ideal

coarse-grained score function would recover the broad features of the all-atom energy landscape while smoothing over the local ruggedness.

The performance of the previous SymDock algorithm of André *et al* is shown in Figures 4.2A–B and D–E, which compare the docking landscapes after the coarse-grained phase and the full protocol for two example proteins, *viz.* Rhamnulose-1-phosphate aldolase (PDBID: 2V9N, symmetry: C4) and snRNP Sm-like protein (1H64, C7). Each model is the end-state of a global docking simulation and is represented as a point in terms of its deviation from the native conformation (root-mean-square deviation of C_α atoms) and its energy predicted by the given score function. The Rosetta all-atom score function[77,78] scores models close to the native conformation more favorably than non-native models (Figures 4.2A and D). However, the energy 'funnels' are absent for SymDock's coarse-grained centroid score function (Figures 1B and E, grey circles). The centroid score function does not score models under 5 Å $RMSD_{C\alpha}$ any better than those far away from the native. Thus, the lowest-scoring structures in the coarse-grained phase are not useful input models for high-resolution refinement.

I considered the characteristics of the centroid score function to help identify opportunities to improve its accuracy. For the centroid score function environment and interacting residue pair terms, only the distance between backbone C_α atoms of two interacting residues across the interface is considered.[26] Previous studies showed that this score function does not sufficiently discriminate near-native interfaces of heterodimers.[177] Also, to prevent favoring non-specific interactions across large, spurious interfaces, the residue-residue contact count of the centroid score function is capped, but this cap hinders the discrimination of large

interfaces. Together, these score function features lead to the flat landscapes and false-positive energy wells observed in Figures 4.2B and E.

Next, I tested whether better discrimination could be obtained by replacing the environment, pair, and contact scores with higher-resolution information about residue backbone orientation. For heterodimeric complexes, we recently developed Motif Dock Score (MDS), which radically improved coarse-grained interface detection.[79] MDS is based on a residue-pair transform framework.[7] It estimates the minimum all-atom score for two residues interacting with a given backbone geometry defined by the six-dimensional transform (three rotations and three translations) required to superimpose the backbone atoms of one residue onto the other. To discretize the transform space, we use 2 Å grids for the translational dimensions and 22.5° grids for the rotational dimensions. For each residue type pair, we have pre-tabulated the lowest observed all-atom scores for every orientation present in high-resolution protein structures in the Protein Data Bank.[117] If the orientation is not observed for that residue type pair (as is the case for the majority of the orientations), we score it as zero. To score a particular conformation in the coarse-grained phase, we look up the residue pair scores from these tables for every residue pair across the interface(s) of the principal subunit and sum them. The symmetry definition is then used to scale the score for the complex.

Figures 4.2C and F (grey points) shows that MDS scores near-native (<5 Å $RMSD_{C\alpha}$) models better than far-away models. The general shape of the MDS energy landscape resembles that of the all-atom score function, with the aforementioned energy funnel near zero $RMSD_{C\alpha}$. Moreover, of the 5,000 coarse-grained models obtained with MDS, 101 and

130 of them have $RMSD_{C\alpha}$ values of less than 2 Å for Rhamnulose-1-phosphate aldolase and snRNP Sm-like protein, respectively, including 86 sub-angstrom models for the former. For comparison, there are no models within 2 Å $RMSD_{C\alpha}$ for the coarse-grained phase with centroid score.

Another comparison of MDS score to the centroid score function is in the ranking of near-native models generated by re-docking the native assemblies. Ideally, the spread of $RMSD_{C\alpha}$ values should be minimal and the models should score better than the global docking models, as they represent the optimal solutions. Starting from the native configuration, centroid score forces the subunits to move away, with median $RMSD_{C\alpha}$ of 3.4 Å and 4.6 Å for Rhamnulose-1-phosphate aldolase and snRNP Sm-like protein, respectively (Figures 4.2B and E, blue points). They also do not score better than the global docking models. In contrast, MDS scores them the lowest with median $RMSD_{C\alpha}$ of 0.6 Å and 0.8 Å for the aforementioned complexes, respectively (Figures 4.2C and F, blue points). Thus, MDS improves the docking performance of the coarse-grained phase, both in terms of the number of near-native models obtained and the ability to discriminate them.

Next, I expanded the comparison to a balanced benchmark of 43 complexes from the two most commonly found symmetry groups, cyclic and dihedral: five each for C2, C3, C4, C5, C6, D2, and D3 symmetries and two each for C7, C8, C9, and D4 symmetries. I challenged the methods with the hardest use-case, *viz.* no information is known apart from the sequence and the point symmetry. This test is akin to a round of the blind docking challenge, Critical Assessment of PRediction of Interactions (CAPRI), where no homologous complex exists for

the modeling target. Starting from a homology-modeled monomer each complex, I generated 5,000 models (see Section 4.5.1 and Appendix Table B.1). In the 5, 50, and 500 top-scoring models, I counted the number of models within 5 Å $RMSD_{C\alpha}$ of the native structure. Table 4.1 compares the bootstrapped averages for the coarse-grained phase run with centroid score and with MDS. For MDS, on average 1.96 of the 5 top-scoring models are near-native compared to 0.32 for centroid. MDS has a superior performance for the 50 and 500 top-scoring models as well.

Table 4.1: Average counts of near-native structures for the 5, 50, and 500 top-scoring models after the coarse-grained phase for coarse-grained score functions.

Score Function	⟨*N5*⟩	⟨*N50*⟩	⟨*N500*⟩
Centroid Score	0.32 ± 0.13	4.2 ± 0.8	25.5 ± 2.7
Motif Dock Score	1.96 ± 0.18	14.9 ± 1.1	41.3 ± 4.0

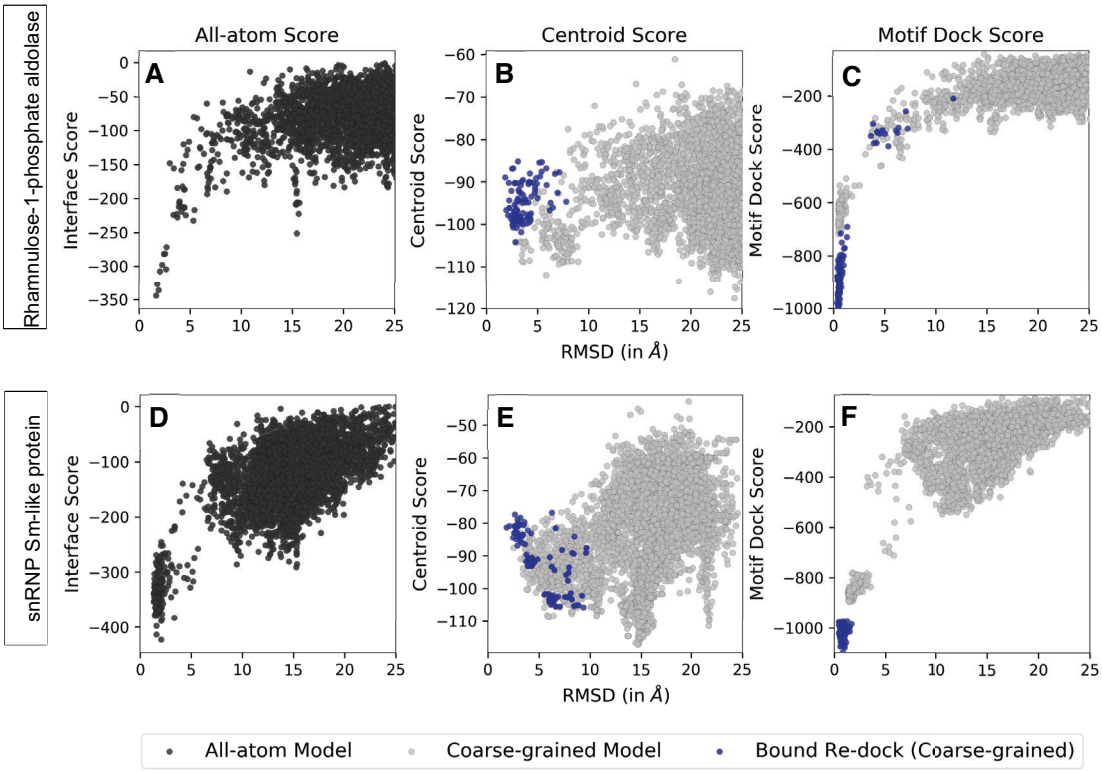


Figure 4.2: Comparison of energy landscapes in all-atom phase and coarse-grained phase. Score versus $RMSD_{C\alpha}$ plots for two representative complexes, Rhamnulose-1-phosphate aldolase (A, B and C) and snRNP Sm-like protein (C, D and E) for 5,000 models generated from global docking of homology-modeled monomers (black, grey circles) and 100 models generated by re-docking of bound subunits (triangles). Coarse-grained energy landscapes with Motif Dock Score (MDS) (C and F) resemble the all-atom energy landscapes (A and D), but those with Centroid Score (B and E) do not. Starting from the homology-modeled monomers, none of the 50 top-scoring models generated using Centroid Score are within 5 Å $RMSD_{C\alpha}$. All the 100 top-scoring models generated using MDS are under 3 Å $RMSD_{C\alpha}$. When re-docking bound subunits, closest models generated using Centroid Score (B and E) have 1.9 Å $RMSD_{C\alpha}$ and high relative scores in both cases. Bound re-docking with MDS (C and F) produces over 80% of the models docked to within 1 Å $RMSD_{C\alpha}$ in both cases. These sub-angstrom re-docked models also score more favorably than all docking models made using homology-modeled monomers. Hence, Centroid Score does not recognize the energy well near the native conformation, whereas MDS does.

4.3.2. Fixed-backbone refinement is insufficient to enter the binding funnel

After the initial arrangement of the subunits was calculated in the coarse-grained phase, I sought to produce a physically-realistic all-atom model. To do so, SymDock reintroduces the side chains, packs interface side chains, and refines the model by fixed-backbone energy minimization while allowing small rigid-body motions. With the coarse-grained phase using MDS producing accurate subunit arrangements, I presumed that this refinement would produce high-quality models. Surprisingly, models that were near-native after the coarse-grained phase had positive (unfavorable) interaction energy scores after refinement, indicating that the docked subunits scores worse than non-interacting monomers. Figure 4.3A shows the MDS binding funnel for *Xenopus* Nucleophosmin (1XB9, C5) with models that have a positive post-refinement interaction score labelled red. About 22% of all structures are unfavorable, and all but one near-native (less than 5 Å $RMSD_{C\alpha}$) are unfavorable. To confirm steric obstruction, I counted clashes as per the CAPRI definition.[178] Figure 4.4 shows that in the 20 lowest-$RMSD_{C\alpha}$ models after fixed-backbone refinement, the average number of inter-chain atom-atom clashes is 50.6, compared to 21 in the native structure. I observed this insufficient refinement of near-native models for most complexes.

To test whether the refinement protocol works with an amenable backbone conformation, I generated 100 all-atom models starting from the native structure of Nucleophosmin. The average number of inter-chain clashes after refinement was 18.7, which was significantly lower

than that of the global docked models (Figure 4.4). All models were under 0.7 Å $RMSD_{C\alpha}$ from the native structure and had highly favorable interaction energies (Figure 4.3B). Since (a) the coarse-grained phase produces near-native subunit arrangements, and (b) fixed-backbone refinement can discover the binding funnel with the right backbone conformation, backbone errors in the homology-modeled monomers were likely causing the clashes in the docked models.

For the global docking simulations, four of the five homology-modeled monomers had backbones under 0.4 Å $RMSD_{C\alpha}$, which was sufficient for assembling the subunits at the coarse-grained level, but insufficient for avoiding steric clashes with the side chains present. (In heterodimer docking, a monomer backbone with RMSD of 0.6 Å is typically sufficient for docking.[62]) I speculated that when symmetry is enforced on an all-atom model, the leeway for backbone variation is markedly reduced. Minor deviations from the native backbone result in substantially higher energies as exemplified in Figure 4.3B where a drop of 117 energy units takes place in 0.25 Å $RMSD_{C\alpha}$.

Compared to heterodimers where the average binding funnel slope is 15 $Å^{-1}$, the slope for this complex was unusually steep. For the homomeric complexes in my benchmark, I found the average slope of the binding funnel to be 249 $Å^{-1}$. Further, the average radius of the binding funnel was found to be 0.26 Å for these complexes as opposed to 0.41 Å for heterodimers (see Section 4.5.10). These observations are conceptually represented in Figures 4.3C and D, where homomers have a narrower, steeper well in the rugged all-atom energy landscape as compared to heterodimers. More examples of binding funnel data for homomers and

heterodimers can be found in Figure 4.5. As homomers generally have extensive interfaces owing to multivalent interactions, I normalized the slopes by dividing them by the lowest interface score observed for the complex to obtain slopes of 0.62 $Å^{-1}$ and 0.31 $Å^{-1}$, respectively for homomeric and heterodimeric complexes. Even after normalization, funnels in homomers are twice as steep as heterodimers. I concluded that in homomers, for a backbone with errors, no amount of side chain packing can help it find the narrow binding funnel. Flexible-backbone strategies are required to reduce steric clashes and build physically-realistic models.

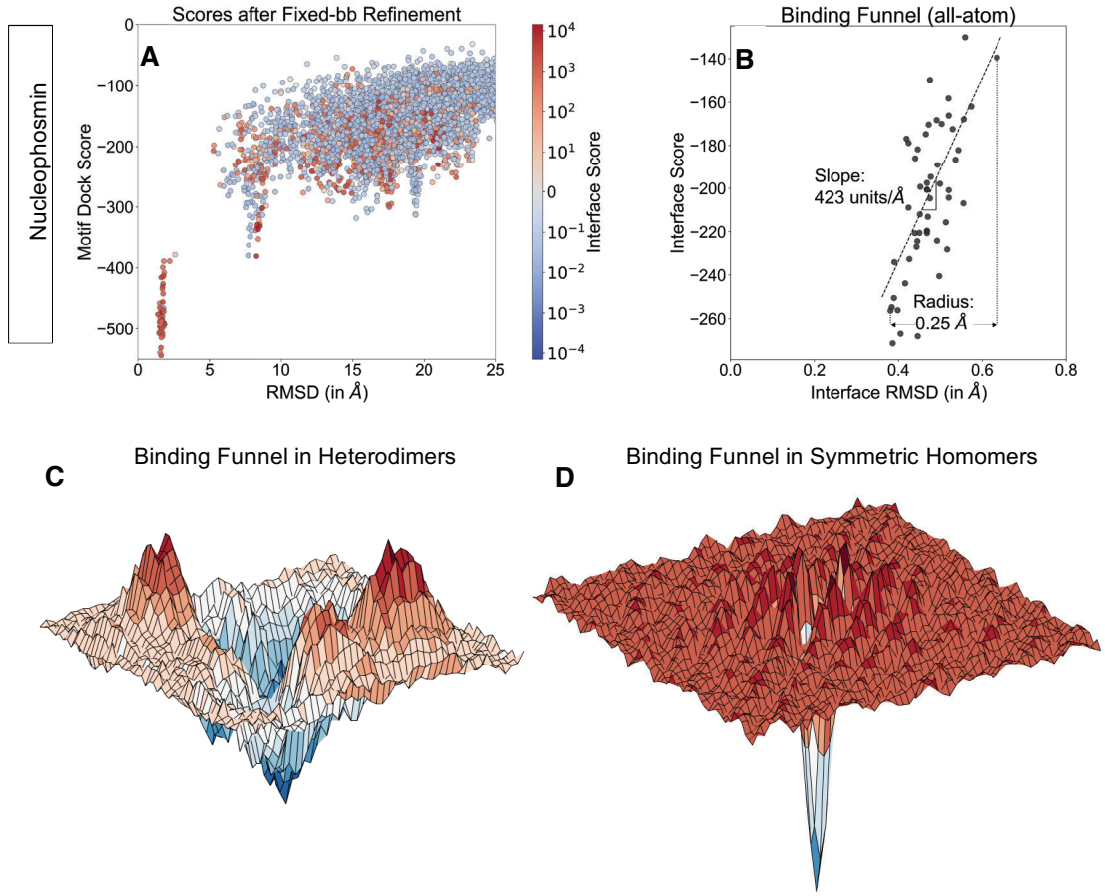


Figure 4.3: Fixed-backbone refinement is insufficient to enter narrow binding funnel. (A) Coarse-grained score versus RMSDCα (after coarse-grained phase) plots for Xenopus Nucleophosmin for 5,000 models. Models are colored by their interface score after fixed-backbone refinement. Almost all models under 5 Å RMSDCα have a positive interface after fixed-backbone refinement arising from minor clashes due to the introduction of side chains, despite repacking. Consequently, these models are discarded. (B) Interface score versus RMSDCα (after full-protocol) plots for Xenopus Nucleophosmin. A rapid drop in interface score between 0.6 and 0.4 Å RMSDCα leads to an energy funnel with steep slope (dashed line) of 423 Å-1 and a radius of 0.25 Å. (C) Conceptual representation of the energy landscape near the binding funnel for heterodimers. The funnel is comparatively shallow with local minima near it. (D) Conceptual representation of the energy landscape near the binding funnel for homodimers as seen by symmetrical docking protocols. The funnel is narrow and steep with no comparable local minima.

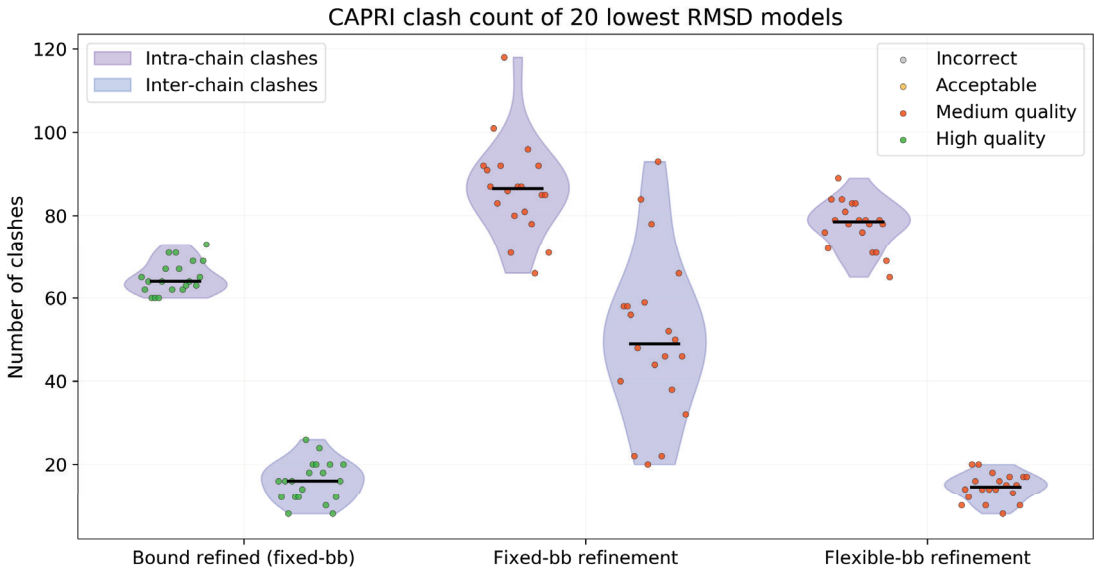


Figure 4.4: Count of intra-chain and inter-chain clashes for interface residues of Xenopus Nucleophosmin as per CAPRI definition. The 20 lowest RMSD models are chosen. Flexible-backbone refinement of complex starting from homology-modeled monomer reduces inter-chain clashes to be close to that observed after fixed-backbone refinement of the native structure.

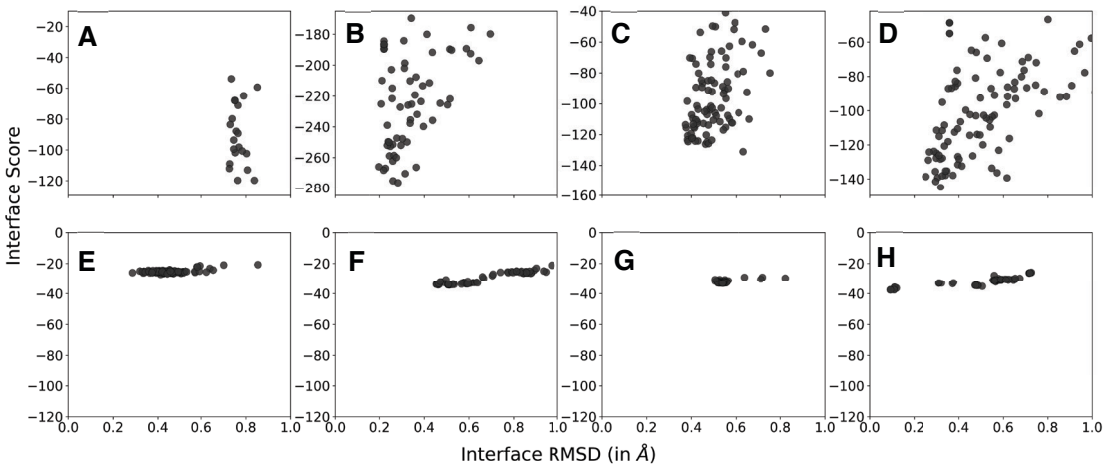


Figure 4.5: Comparison of interface score versus $RMSD_{C\alpha}$ plots produced by native refinement of homomers and heterodimers. The four example homomers are: (A) 3,2-trans-enoyl-CoA isomerase (1SG4, C3), (B) snRNP Sm-like protein (1H64, C7), (C) gp23.1 chaperone (2XF7, C6), and (D) Cytolysin (4OWK, C7). The four example hetero-dimers are: (E) APR-APRin complex (1JIW), (F) L. casei HprK/P - B. subtilis HPr (1KKL), (G) Glutamyl-tRNA synthetase (2HRK), and (H) IL-13 and C836 FAB (3L5W). In all plots, the y-axis spans 120 energy units. In general, homomer binding funnels are deeper, steeper and narrower.

4.3.3. In context, flexible backbone refinement is crucial to enter the binding funnel

To find alternative routes to enter the binding funnel, I tried mimicking natural mechanisms of backbone flexibility. Two kinetic mechanisms widely observed in assembly and regulation of proteins are conformational selection and induced fit.[95]

4.3.3.1. Imitating conformational selection

We have previously leveraged conformational selection to improve the docking performance of heterodimeric complexes by pre-generating an ensemble of backbone

conformations from the individual monomers and docking the optimal backbones.[79] Using a similar approach, starting from a homology-modeled monomer, I generated 50 conformers each using three conformer generation methods: perturbations along the normal modes by 1 Å,[179] small backbone perturbations using Backrub,[108] and general refinement using Rosetta's Relax protocol[107] (see Section 4.5.3). I supplemented the ensemble of the five original homology-modeled backbones with the new 150 backbone conformations. I ran 500 independent fixed-backbone simulations with each of the 155 monomer backbones and bootstrapped the results to simulate the selection of 2,500 models (see Section 4.5.9).

Next, I tested the efficacy of starting with these large, diverse ensembles using a small benchmark of 10 cyclic complexes. I compared the number of structures with $RMSD_{C\alpha}$ less than 5 Å from the native in the 1% top-scoring models, i.e. the 25 top-scoring models. Figure 4.6 shows a case-by-case comparison. Docking with just the homology models (HM/Fixed-bb) gives a median value of 9.6 near-native models after the coarse-grained phase, which goes down to 3.0 after the full protocol. Using a mixture of conformations (HM+Ens/Fixed-bb), the results get marginally worse with median values of 6.8 and 2.8 near-native models, respectively, for the coarse-grained phase and the full protocol. Starting with a large ensemble improves performance for some complexes and makes it worse for others. In general, backbone conformations generated from the monomer lack information about where the other subunits are and encounter the same barriers as the original homology models. Thus, my conformational selection approach was unable to improve docking accuracy for symmetric homomers.

4.3.3.2. Imitating induced fit

I next hypothesized that the backbone needed to be adjusted in the context of the complex and not independent of it. That is, since the coarse-grained phase was correctly predicting the rigid-body arrangement of the subunits, I tested using these coordinates to induce a backbone fit at the interface. Specifically, I alternately repacked side chains and minimized the energy of the whole protein while slowly ramping up the repulsive component of the van der Waals potential from 2% to 100%. Gradient-based energy minimization along the backbone dihedral angles (φ and ψ) provided an avenue for the backbones to relieve clashes. The presence of the other subunits provided the necessary constraints to move the backbone to best fit the complex. To ensure a constant context, I removed rigid-body moves. Starting with just the five homology-modeled monomers per complex, I generated 5,000 docked models.

Finally, I tested this approach for the same benchmark of ten proteins (HM/Flexible-bb) and bootstrapped the results to simulate the selection of 2,500 models (section 4.5.11). In the top-scoring 1% of models, the median counts of near-native models increased to 21.1 after the coarse-grained phase and 22.3 after the full protocol (Figure 4.6). Thus, I conclude that inducing a change in the backbone retains good coarse-grained models and gains additional near-native models for all complexes tested. Further, the average number of inter-chain clashes in the 20 lowest-$RMSD_{C\alpha}$ models decreases from 50.6 in fixed-backbone refinement to 14.5 (Figure 4.4).

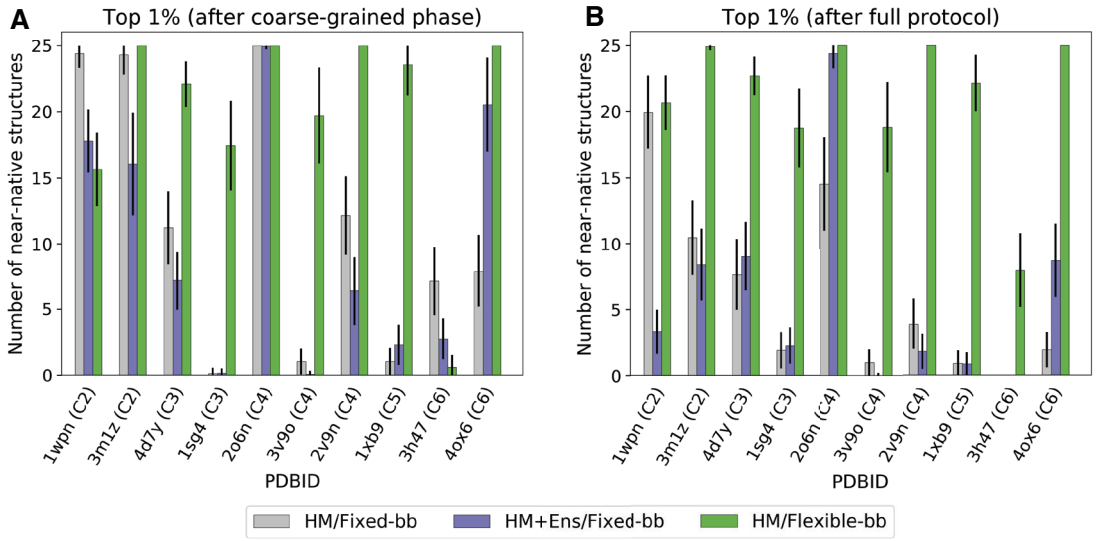


Figure 4.6: Flexible-backbone refinement improves docking performance. Comparison of bootstrapped averages of the number of near-native structures in the set of 2,500 docking models using: [grey] the homology models (HM) and fixed-backbone refinement, [blue] homology models supplemented with an ensemble of 150 pre-generated backbone conformations (HM+Ens) and [green] fixed-backbone refinement, and the homology models and flexible-backbone refinement after the coarse-grained phase (A) and after the full protocol (B). Starting with 150 additional backbone conformations generated without the context of the complex improves docking performance for 4 out of 10 complexes, but makes it worse for 5 complexes. Starting with just the homology models and performing flexible-backbone refinement leads to improvements in 9 out of 10 complexes after the coarse-grained phase and in all complexes after the full protocol. After flexible-backbone refinement, more than 70% of the top-scoring models were near-native for 9 out of the 10 complexes.

4.3.4. Improvement in global docking performance over a diverse benchmark

In the two-stage Rosetta SymDock2 protocol (Figure 4.7), I combine the coarse-grained phase with MDS with an in-context, flexible-backbone, all-atom refinement. To evaluate the performance of Rosetta SymDock2 and compare it to SymDock across a benchmark of 43

proteins, I performed a global docking search along symmetrical conformation space starting from five homology-modeled input monomers per target to generate 5,000 candidate models for each complex. Next, I resampled the docked models and reported averages and medians for targets success metrics based on the near-native model counts. For the coarse-grained phase, I defined a near-native model as one with $RMSD_{C\alpha}$ under 5 Å. For the full protocol, I defined near-native as acceptable, medium-quality, or high-quality as per the CAPRI criteria, which are based on the ligand $RMSD_{bb}$, interface $RMSD_{bb}$, and fraction of native contacts recovered[178] and detailed in Section 4.5.12.

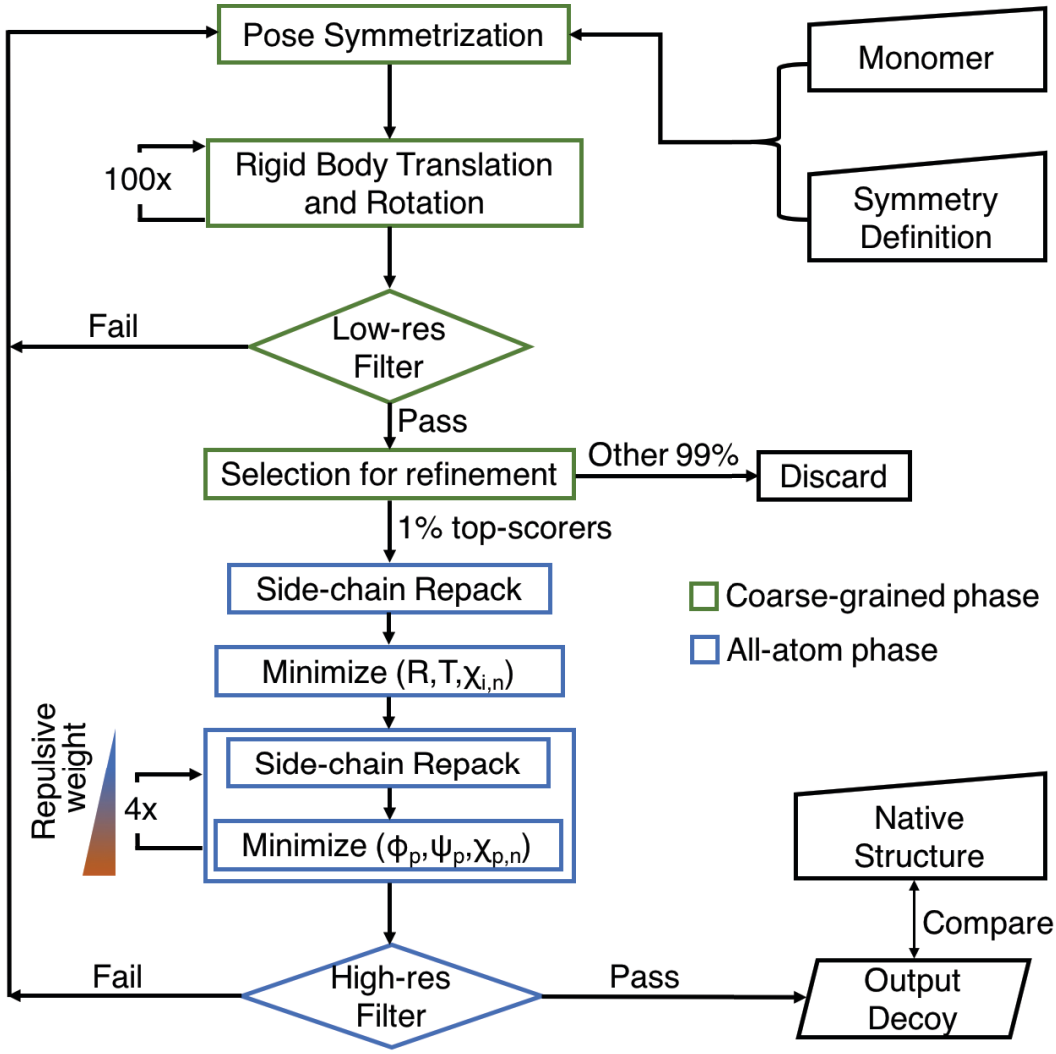


Figure 4.7: Flowchart describing major steps in Rosetta SymDock 2 protocol. In the all-atom phase the structure is initially minimized along rotational rigid body coordinates (R), translational rigid body coordinates (T), and the dihedrals of the interface residue side chains ($\chi_{i,n}$, where $i \in$ interface). This is followed by four cycles of side-chain repacking and minimization along the dihedrals of all residues ($\{\phi_p,\psi_p,\chi_{p,n}\}$, where $p \in$ protein). Each cycle is carried out at a different weight of the van der Waals repulsive term starting from 2% of the original weight and ramping up to 100%. See Methods for filter descriptions.

To test near-native sampling and discrimination ability, I counted the number of near-native models in the five top-scoring models and averaged over resampling attempts to calculate the ⟨*N5*⟩ metric (Section 4.5.11). ⟨*N5*⟩ after the coarse-grained phase indicates the ability of the broad search in the coarse-grained space to find approximate solutions. Most importantly, ⟨*N5*⟩ after the full protocol determines the overall accuracy of the method. For SymDock2, the average ⟨*N5*⟩ value improved from 2.0 to 2.8 going from the coarse-grained phase to the full protocol, indicating that while the broad search in the coarse-grained space found approximate solutions, the introduction of side chains and flexible-backbone refinement further discriminated near-native models. SymDock had an average ⟨*N5*⟩ value of 0.3 for both the coarse-grained phase and 0.8 for the full protocol suggesting a failure to sample and discriminate near-native models in most complexes. Figure 4.8 presents a case-by-case comparison between the two methods, and Table 4.2 provides a category-wise summary of the benchmark results. The average performance on cyclic complexes is better than on dihedral complexes on every metric. Detailed metrics and plots for each complex can be found in Appendix Table B.3 and Figures B1–B4.

I classified a homomeric complex as successfully docked if ⟨*N5*⟩ ≥ 3, i.e. at least 3 of the 5 top-scoring models are near-native on average. This criterion indicates that the protocol converges on the native structure. While SymDock docked 4 of the 43 complexes successfully, SymDock2 docked 24 of them successfully, representing a six-fold improvement in the success rate of blind docking for a general case. I observed performance gains for both symmetry

groups, with 15 new cyclic complexes and 5 new dihedral complexes being docked successfully.

To estimate how many independent trajectories must be run to completion, I evaluated the fold-enrichment of near-native models for the top-scoring 1% of models after the coarse-grained phase, $\langle E_{1\%} \rangle$. SymDock2 had an average $\langle E_{1\%} \rangle$ value of 29.3, indicating a highly enriched low-scoring model set, while SymDock had a lower average $\langle E_{1\%} \rangle$ value of 6.6. Thus, if I were to only refine the top-scoring 1% of models after the coarse-grained phase of SymDock2, the average $\langle N5 \rangle$ value after the full protocol would be 2.9. Furthermore, the number of complexes successfully docked for SymDock2 increases to 25. To explain the increase in success rates, I consider the example of Acylhomoserine lactonase (4ZO2, C2). With the all-atom score function, a non-native binding mode around 10 Å $RMSD_{C\alpha}$ from the native is scored more favorably than near-native conformations (Appendix Figure B.3). MDS discriminates the native biding mode better and no models having the aforementioned non-native binding mode are selected in the top 1% (Appendix Figure B.1). By reducing the number of false positives, the success rates are increased. Thus, I recommend running SymDock2 as a two-step protocol where only the top-scoring 1% of coarse-grained models are refined.

Table 4.2: Category-wise summary of the results of Rosetta SymDock and SymDock2 across a benchmark of 43 complexes.

	Category	Rosetta SymDock			Rosetta SymDock2		
		Coarse-grained $\langle E_{1\%} \rangle$	Coarse-grained $\langle N5 \rangle$	Full protocol $\langle N5 \rangle$	Coarse-grained $\langle E_{1\%} \rangle$	Coarse-grained $\langle N5 \rangle$	Full protocol $\langle N5 \rangle$
Average Value [a]	**Cyclic** *(n = 31)*	6.7	0.4	1.1	34.6	2.6	3.1
	Dihedral *(n = 12)*	6.4	0.1	0.2	15.8	0.2	1.8
	All *(n = 43)*	6.6	0.3	0.8	29.3	2.0	2.8
	All w/ 1% filter [c]	N/A	N/A	1.0	N/A	N/A	2.9
Expected Success [b]	**Cyclic** *(n = 31)*	6	2	4	14	16	19
	Dihedral *(n = 12)*	0	0	0	0	0	5
	All *(n = 43)*	6	2	4	14	16	24
	All w/ 1% filter [c]	N/A	N/A	4	N/A	N/A	25

[a] Values are the average of bootstrapped means across all complexes of the specified category.

[b] Expected success is the number of successfully-docked complexes based on the following criteria: for *N5*, $\langle N5 \rangle \geq 3$; for $E_{1\%}$, $\langle N50 \rangle \geq 15$.

[c] Only the 1% top-scoring models after the coarse-grained phase underwent all-atom refinement.

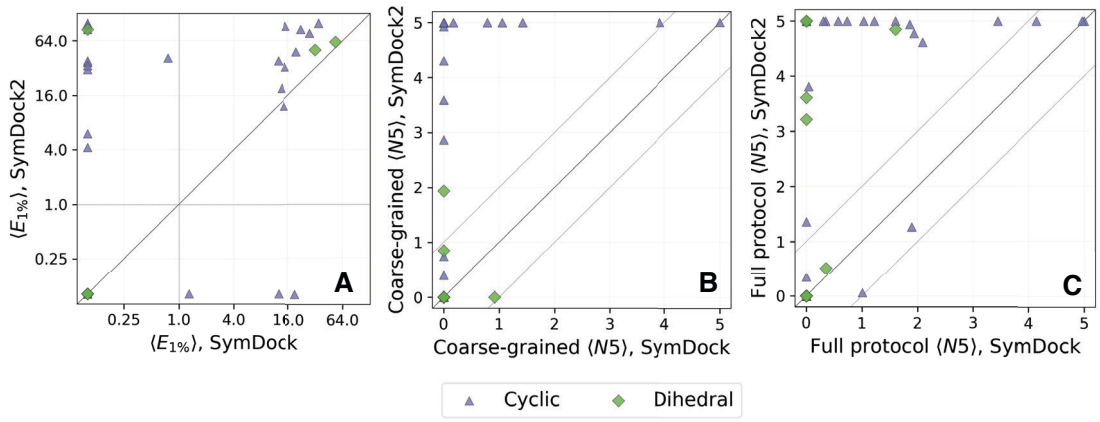


Figure 4.8: Rosetta SymDock2 compares favorably with SymDock on various assessment metrics. Comparison of bootstrapped-averaged metrics for 43 individual complexes (31 cyclic complexes [triangle] and 12 dihedral complexes [diamond]) both after the coarse-grained phase (A and B) and after the full protocol (C) shows significant performance gains. All complexes (points) above the diagonal line are improved in SymDock2. (A) Comparison of fold-enrichment of near-native models in the 1% top-scoring models, $\langle E_{1\%} \rangle$ on a log-log scale shows a higher enrichment in 19 cyclics and 3 dihedrals and a lower value in 4 cyclics and 0 dihedrals. Complexes to the right of the vertical dashed line are enriched in SymDock, and complexes above the horizontal dashed line are enriched in SymDock2. (B and C) Comparison of number of near-native models in the five top-scoring models, $\langle N5 \rangle$, shows marked improvements both after the coarse-grained phase (B) and after the full protocol (C). Areas above and below the dashed lines indicate cases where the two methods differ significantly, *i.e.* by more than 1 model on average. SymDock2 has significant improvements in 16 cyclics and 1 dihedral complex after the coarse-grained phase, and most importantly, in 17 cyclics and 5 dihedrals after the full protocol. No complexes were modeled significantly worse with SymDock2.

4.3.5. Flexible-backbone refinement does not affect net efficiency

Compared to fixed-backbone refinement, modeling backbone motions requires sampling an exponentially larger conformational space. Instead of explicitly sampling backbone changes, I employed systematic energy minimization along backbone torsions to induce a fit. In fixed-

backbone refinement of the interfaces, the computational time depends on the interface size and is largely independent of the monomer size. On the other hand, energy minimization along the backbone involves small changes in the subunit core to better accommodate the interfaces and hence, the time increases with monomer size. In fact, SymDock2 was between 2 and 3 times slower than SymDock for models that had larger interfaces to be fit (data not shown). Flexible backbone refinement also led to the fitting of spurious interfaces that were then weeded out by their relatively poor interface scores. As a result, almost every SymDock2 model had a negative interface score. Compared to SymDock, where a significant number of models are filtered out because of positive interface scores, induced fit reduces the total number of models rejected with the same filters (see Section 4.5.8 for filter values). As a result of the low rejection rate, SymDock2 can compensate for the additional time required for flexible-backbone refinement by attempting fewer trajectories. We have previously shown that MDS is marginally faster than centroid score in the coarse-grained phase,[79] which too works in favor of SymDock2. For an even comparison, in Figure 4.9, I show the time per model for the two methods for every complex when all coarse-grained models are refined. SymDock was faster for 22 complexes and SymDock2 for 20 complexes. SymDock was typically faster for larger complexes and SymDock2 for smaller complexes. In 31 of the 43 complexes the run time difference was less than ± 20%, with the largest difference being less than 70%.

High fold-enrichment of near-native models for the 1% top-scoring models after the coarse-grained phase allowed us to considerably reduce the number of models refined using the expensive all-atom refinement. Broad sampling in the coarse-grained phase takes 51–78%

of the time in each trajectory. By carrying forward only the top 1% from the coarse-grained phase to the refinement phase, one could save 22–49% of the total time. For the average complex in my benchmark, to generate 5,000 coarse-grained models and then refine 50 of them, SymDock2 requires 89 hours on a 4-core personal computer or under 1 hour on a 360-core cluster.

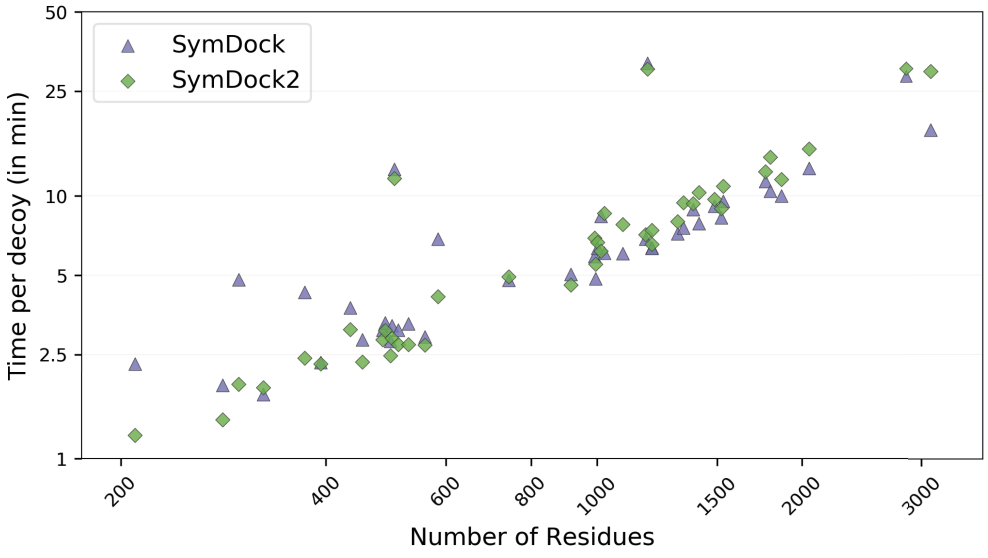


Figure 4.9: On average, Rosetta SymDock and SymDock2 have similar per-decoy runtimes in the benchmark. Comparison of average time per decoy on a log-log plot demonstrates similar scaling with complex size and symmetry for SymDock (triangle) and SymDock2 (diamond). Despite having a slower all-atom refinement phase, no complex had a more than a 70% overhead with SymDock2. For the two methods, run times were within ± 20% for 31 out of 43 complexes.

4.4. Discussion

Here, I have developed and benchmarked a method to accurately model homomeric assemblies from an approximate monomer structure and the point symmetry.The first innovation was using a six-dimensional coarse-grained scoring scheme, MDS, to successfully discriminate near-native interfaces with accuracy comparable to an all-atom score function. The second innovation was refining approximate models with small backbone motions to fit tight complexes together. Taken together, these two advances achieve successful blind global docking of six times as many complexes as Rosetta's original SymDock.[67] In Table 4.3, I compare the global docking performance of Rosetta SymDock and the new SymDock2 to four leading homomer docking methods recently tested by Yan *et al*: SymmDock,[65] M-ZDOCK,[66] SAM,[180] and HSYMDOCK.[69] In this table, to compare to the other methods, I changed my success criterion to match their criterion of $\langle N10 \rangle \geq 1$, *i.e.* at least one of the ten top-scoring models should be CAPRI acceptable, medium- or high-quality. While the methods are tested on different benchmarks with different ways of generating unbound structures, general patterns can be observed. With a success rate of 71% for cyclic complexes, Rosetta SymDock2 outperforms other methods. For dihedral complexes, SymDock2's success rate of 50% is comparable to HSYMDOCK. Moreover, flexible refinement in SymDock2 ensures that the interfaces are relatively free of clashes, which were frequently observed with fixed-backbone docking of tightly-packed homomers.

I also explored some characteristics of the interfaces of homomers and compared them to those of heterodimers. We developed MDS with the conjecture that given the relative orientation between backbones of interacting residues, we can estimate the optimal side chain interaction energy. The broad bin size of the score tables prevents overfitting to any particular protein class.[79] Although MDS was developed to recognize hetero-oligomeric interfaces, it performs just as well for homomers, which suggests that at the level of individual residue pairs, the interfaces of homo- and hetero-mers have similar interactions.

I found that the near-native binding energy landscape was 14 times steeper on average than that of heterodimers. Even after normalizing for the depth of the binding funnels, the homomer funnels were twice as steep as those of heterodimers. This energy landscape is seen by symmetric docking protocols with enforced symmetry, which is not a constraint for the natural association of the subunits for symmetries higher than C2. For example, D2 complexes likely assemble as dimer of dimers with the ratio of the interaction strengths of the different interfaces dictating the hierarchy of assembly.[181] For some proteins, inter-subunit interactions may be essential to find an energy funnel while folding,[163] and hence an independent docking landscape may not exist. Reproducing these multi-state interactions becomes infeasible for a general case where the pathway of association is unknown, and so we resort to docking all the subunits together symmetrically. However, once assembled, the depth of the energy funnel suggests that symmetry confers stability through multivalent associations.

Most of the proteins that I have considered in the benchmark are globular, which allowed us to deconstruct the problem into generating an approximate monomer and then docking

and refining it. The homology server I used, Robetta, performed admirably with monomer modeling. For the few cases for which it did not produce a monomer model under 2 Å RMSD, my docking performance suffered. For example, for phage SF6 terminase small subunit (3ZQO), the complex is stabilized by intertwined interfaces, and hence Robetta failed to create a good monomer model. Such proteins require simultaneous folding and docking. Previous fold-and-dock attempts have achieved success rates similar to that of symmetrically docking small globular proteins.[70] However, owing to the sheer size of the conformational space that needs to be sampled, without experimental constraints, *de novo* folding and docking is currently feasible only when subunits are smaller than 100 residues. Instead, incorporating symmetry information while homology modeling provides a promising avenue, which servers like Robetta,[182] SWISS-MODEL Oligo,[183] and GalaxyHomomer[68] have recently demonstrated. Unfortunately, >90% sequence identity is required to guarantee symmetry type conservation;[71] at <30% identity, interactions may differ completely.[17] In case of Robetta, 11 of the 22 complexes of the CASP12 experiment[184] did not have sufficient symmetric templates and in case of SWISS-MODEL Oligo, 20% of the complexes considered did not have any viable symmetric templates. Thus, for large homomers, especially for those without close homologs, symmetric docking methods are required for modeling the complex.

The versatility of the techniques developed here facilitates application across a broad spectrum of problems. Integration with Rosetta's input system allows us to incorporate cryo-EM, NMR, SAXS, cross-linking, and sequence co-evolution data.[173–176] In combination with these data, SymDock2 is a powerful tool for understanding homomer assembly and function.

Table 4.3: Comparison of leading symmetrical homomer docking methods with Rosetta SymDock2.

Method	Description	Accuracy	
		Cyclic Complexes	Dihedral Complexes
SymmDock (Schneidman-Duhovny et al., 2005)[65,a]	Local feature matching, cluster evaluation	79/213 (37.1%)	N.A.[d]
M-ZDOCK (Pierce et al., 2005)[66,a]	FFT docking, model evaluation	86/213 (40.4%)	N.A.[d]
SAM (Ritchie and Grudinin, 2016)[180,a]	Spherical polar FFT docking, model evaluation	94/213 (44.1%)	13/35 (37.1%)
HSYMDOCK (Yan et al., 2018)[69,a]	FFT docking, cluster evaluation	104/213 (48.8%)	19/35 (54.3%)
Rosetta SymDock (André et al., 2007)[67b,c]	Monte Carlo docking, model evaluation	15/31 (48.4%)	2/12 (16.7%)
Rosetta SymDock2 (This article)[b,c]	Monte Carlo docking, model evaluation	22/31 (71.0%)	6/12 (50.0%)

[a] Benchmark set: Yan et al., 2018[69]
[b] Benchmark set: this article
[c] For an even comparison of all the methods, the metric for success was changed from $\langle N5 \rangle \geq 3$ to $\langle N10 \rangle \geq 1$.
[d] Dihedral complex docking is not available with this method.

4.5. Methods

4.5.1. Benchmark set generation

I generated a benchmark of symmetric homomeric proteins from structures deposited in the Protein Data Bank (PDB) having either cyclic (C2–C9) or dihedral (D2–D4) point symmetries. First, I filtered structures by resolution, retaining those with a resolution of 1.5 Å or better for C2–C4 and D2–D4, with a resolution of 2.0 Å or better for C5–C7, and with a resolution of 2.5 Å or better for C8–C9. I then randomly chose complexes from each symmetry group and retained those that passed the following selection criteria. I discarded entries with atoms having B factors greater than 40 or ligands with more than 5 atoms at the interfaces. Additionally, I only included entries for which the biologically relevant symmetry is confirmed in a publication. Next, I discarded any entry for which the earliest `REVDAT` record date was earlier than 2002. Within each symmetry group, I selected complexes to include a range of monomer sizes and diversity in secondary structural elements. I did not filter out complexes with intertwined interfaces. In total, I selected a benchmark of 43 complexes of different symmetries: five each for C2, C3, C4, C5, C6, D2, and D3 symmetries and two each for C7, C8, C9, and D4 symmetries, which are listed in Appendix Table B.1.

From the set of available structures, I randomly chose 10 cyclic complexes for testing flexible-backbone strategies. These complexes are listed with * after the PDB ID in Appendix Table B.1. The full benchmark could not be used as ensemble generation and global docking starting with 150+ monomer conformations were extremely resource consuming.

4.5.2. Generation of homology-modeled monomers

For each protein, I obtained five homology-modeled monomers from the Robetta server.[133] I submitted the FASTA sequence for the first monomer chain to Robetta. To best simulate CAPRI conditions, I instructed Robetta to only consider templates older than the first `REVDAT` record for the complex being modeled, *i.e.* only those templates that would have existed before the PDB was deposited. Secondly, I instructed Robetta not to consider any symmetry information while modeling the monomers. The monomers obtained were used as input structures for the SymDock2 docking protocol.

4.5.3. Generation of alternative conformations from the monomer

From the five monomer models produced for each target by Robetta, the model with the highest backbone RMSD less than 1.5 Å was selected as input for three alternative conformer generation methods in order to sample a variety of backbone conformations. This RMSD cutoff was chosen so that the conformations are not too close to the native structure and not so different that they do not fit in with other subunits. Relax, Backrub, and normal mode analysis (NMA) with perturbation steps of 1 Å were each used to produce 50 structures, resulting in an ensemble of 150 structures per target. The conformations in these ensembles were pooled with the homology-modeled monomers and used as input structures for the SymDock2 docking protocol.

4.5.3.1. Relax

Rosetta Relax is a refinement protocol in which the protein is perturbed using a series of small backbone torsion moves,[107] which is followed by side-chain repacking and energy minimization along all torsion angles (φ, ψ and χ_i). Each perturbation step is carried out at a particular weight of the van der Waals repulsive component of the all-atom score function (`fa_rep`). In each cycle, the weight is progressively ramped from 20% of the maximum (to allow atoms to come closer) to 100% (to resolve clashes). The lowest energy structure after 20 such cycles is chosen as the final structure for that trajectory. Relax was used to generate 50 monomer conformations.

The Relax protocol was implemented using the following command:

```
relax.linuxgccrelease
-in:file:s <PDB> -nstruct 50 -relax:thorough
```

4.5.3.2. Backrub

Rosetta Backrub attempts to capture small conformational changes that proteins undergo in solution.[108] The protein backbone is divided into segments and each segment is rotated about the axis joining the first and the last backbone atom of the segment, while fixing the rest of the protein. The rotational movements are along six internal backbone degrees of freedom: the φ, ψ, and the N-C_α-C bond angles at each pivot. This is followed by side-chain repacking and energy minimization along all torsion angles (φ, ψ and χ_i). This process is

repeated for 20,000 trials and the lowest energy structure is chosen. Backrub was used to generate 50 monomer conformations.

The Backrub protocol in Rosetta was implemented using the following command:

```
backrub.linuxgccrelease
-in:file:s <PDB>    -backrub:mc_kt 0.6
-nstruct 50 -backrub:ntrials 20000
```

4.5.3.3. Perturbation along normal modes

A normal mode of the protein is a collective motion in which all bonds are vibrating with the same frequency and phase: normal mode analysis (NMA) reveals accessible low-frequency vibrational modes that are thought to capture biologically relevant protein motions.[179] In Rosetta, NMA is implemented via the XML interface RosettaScripts.[118] To generate monomer conformations, the protein is perturbed by steps of 1 Å randomly distributed over the first five normal modes. As this motion disrupts bond angles and bond lengths, the perturbed structure is subsequently relaxed using the aforementioned method with the exception of energy minimization being along Cartesian coordinates of the atoms. The score function of Relax is biased to favor ideal bond angles and bond lengths. This method was used to generate 50 monomer conformations.

The following command was used for running NMA-Relax in Rosetta:

```
rosettascripts.linuxgccrelease
-in:file:s <PDB> -nstruct 50 -parser:protocol nma.xml
```

In this command, `nma.xml` contains the details of the protocol which is outlined below:

```
<ROSETTASCRIPTS>
    <SCOREFXNS>
       <ScoreFunction  name="ref_cart"
weights="ref2015_cart" />
    </SCOREFXNS>
    <RESIDUE_SELECTORS>
    </RESIDUE_SELECTORS>
    <TASKOPERATIONS>
    </TASKOPERATIONS>
<FILTERS>
    </FILTERS>
    <MOVERS>
        <NormalModeRelax    name="nma"    cartesian="true"
        centroid="false"     scorefxn="ref_cart"nmodes="5"
        mix_modes="true"                    pertscale="1.0"
        randomselect="false"
        relaxmode="relax"                       nsample="20"
        cartesian_minimize="false"
        />
    </MOVERS>
    <APPLY_TO_POSE>
    </APPLY_TO_POSE>
    <PROTOCOLS>
        <Add mover="nma" />
    </PROTOCOLS>
    <OUTPUT scorefxn="ref_cart" />
</ROSETTASCRIPTS>
```

4.5.4. Symmetry definitions

I symmetrized the monomeric input structure using symmetry definitions in Rosetta's symmetry framework.[172] I used two kinds of symmetry definitions: general (also called *de novo*), and specific. Symmetry definitions contain information about the rigid-body arrangement of the subunits, how to scale the energy from calculations on one subunit (or a set of subunits), and specification of the degrees of freedom for the system.

I generated general or de novo symmetry definitions for each point symmetry in the benchmark using the pre-packaged Rosetta script, `make_symmdef_file_denovo.py`. This script inputs the symmetry group (Cn or Dn) along with the number of subunits. No information about any specific PDB is supplied. I used these definitions for global docking simulations. For example, the following command generates a general C2 symmetry definition:

```
make_symmdef_file_denovo.py -symm_type cn -nsub 2 > C2.symm
```

I generated specific symmetry definitions for each complex using the pre-packaged Rosetta script, `make_symmdef_file.pl`. This script inputs a symmetric PDB along with chain ID's of the principal subunit (`A`), an interacting subunit (`B`), and in case of dihedrals, the chain ID of a chain in the sub-system in which `A` is not present (`X`). I also specified that I was using non-crystallographic symmetry (`NCS`) and the farthest interacting subunits were at a distance of `d` Å. I used these definitions for local docking, bound re-docking and bound refinement. The following command generates a specific symmetry definition from `<PDB>`:

```
make_symmdef_file.pl –m NCS –p <PDB> –a A -i B X –r d -f >
<PDB>.symm
```

The symmetry definitions were used as inputs for the SymDock and SymDock2 protocols.

4.5.5. Global docking simulations

I performed global docking using general symmetry definitions for the point symmetry of the given complex. The five homology-modeled monomers were used as inputs, each of which was used to start 1,000 independent trajectories to generate a total of 5,000 models. To evaluate backbone flexibility using conformational selection, the five monomer conformations were supplemented with another 150 conformations generated using Relax, Backrub and NMA. For these simulations, each input conformation was used to start 500 trajectories, thus generating a total of 77,500 models.

Global docking simulations with SymDock/SymDock2 used the following command:

```
SymDock.linuxgccrelease
-in:file:l <list_of_input_PDBs> -nstruct n (where n = 500 or
1000)
-symmetry:symmetry_definition <general_symm_def_file>
-symmetry:initialize_rigid_body_dofs
-symmetry:symmetric_rmsd
-ex1   -ex2aro -out:file:fullatom
```

To run just the coarse-grained phase, I removed

```
-ex1   -ex2aro   -out:file:fullatom
```

In SymDock2, the following options were added to enable Motif Dock Score:

```
-docking_low_res_score motif_dock_score
     -mh:path:scores_BB_BB <Path to MDS tables>
     -mh:score:use_ss1 false     -mh:score:use_ss2 false
     -mh:score:use_aa1 true -mh:score:use_aa2 true
```

4.5.6. Bound re-docking

In order to assess the ability of the coarse-grained phase in SymDock and SymDock2 to correctly identify and score near-native subunit arrangement, I re-docked the bound conformation. I started with the native monomer and specific symmetry definition taken from the native PDB, and ran the coarse-grained phase using the following command:

```
SymDock.linuxgccrelease
-in:file:s <PDB_with_native_chain_A>  -nstruct 100
-symmetry:symmetry_definition <specific_symm_def_file>
-symmetry:initialize_rigid_body_dofs
-symmetry:symmetric_rmsd
```

In SymDock2, the following options were added to enable Motif Dock Score:

```
-docking_low_res_score motif_dock_score
     -mh:path:scores_BB_BB <Path to MDS tables>
     -mh:score:use_ss1 false     -mh:score:use_ss2 false
     -mh:score:use_aa1 true -mh:score:use_aa2 true
```

4.5.7. Bound refinement

In order to assess the shape of the energy landscape near the bound conformation, I started with the native structure and perturbed it by iteratively re-packing side chains and minimizing energy along torsion angles at the interface and along the inter-subunit distance. I started with the native monomer and specific symmetry definition taken from the native PDB, and ran only the all-atom refinement protocol using the following command:

```
SymDock.linuxgccrelease
-in:file:s <PDB_with_native_chain_A>  -nstruct 100
-symmetry:symmetry_definition <specific_symm_def_file>
-symmetry:initialize_rigid_body_dofs
-symmetry:symmetric_rmsd
-docking:docking_local_refine
-ex1    -ex2aro -out:file:fullatom
```

4.5.8. Filtering docking models

Both Rosetta SymDock and SymDock2 filter out demonstrably poor models after the coarse-grained stage and after refinement (see Figures S1 and S2). The low-resolution filter after the coarse-grained phase in SymDock is based on terms of the centroid score function. This filter is `interchain_vdw` (penalizes clashes across chains) ≤ 1, `interchain_contact` (penalizes small interfaces) ≤ 10, and `atom_pair_constraint` (penalizes deviations from constraints) ≤ 1. For SymDock2, the low-resolution filter is based on MDS and is `interchain_vdw` ≤ 5. The high-

resolution filter after all-atom refinement is more general and is common to SymDock and SymDock2. It is `total_score` ≤ 1,000,000 (total score of the model should not be ridiculously high) and `I_sc` ≤ 0 (it should be more favorable for the monomers to interact than remain separate).

4.5.9. Simulation of conformational selection and induced fit

Once the ensembles are generated, in heterodimers, by superimposing different backbone conformations of a partner onto the current backbone along the interface, RosettaDock simultaneously samples rigid body orientations and backbone conformations.[79] In SymDock, owing to multiple independent interfaces in homomers, this simultaneous sampling is not feasible since a conformation aligned with one interface may have significant clashes with the other interfaces. Thus, instead of sampling backbones during docking, I ran independent fixed-backbone simulations with each of the 155 monomer backbones. By creating 500 docked models per backbone, I generated a total of 77,500 models per complex. As the total number of docked models was different when using just the homology-modeled monomers and when supplementing it with the ensemble, I simulated the selection of 2,500 models for analysis. (If I had only generated 500 models starting from the five homology-modeled monomers, I would have obtained 2,500 docked models; hence, this value.) Unlike conformational selection, simulating induced fit does not require a multitude of backbone conformations to be sampled independently. Starting with just the five homology-modeled monomers, I generated 5,000 models. For an even comparison, I simulated the selection of 2,500 models for analysis. Thus,

not only does inducing a fit after the coarse-grained phase improve the docking performance, it requires far fewer models than conformational selection to capture relevant backbone motion.

4.5.10. Binding energy funnel characterization

I compared the characteristics of the binding energy funnel in the 43 homomeric complexes in this study with 87 hetero-dimers previously studied.[79] After fixed-backbone refinement of the native structure, the slope of the binding funnel is defined as the slope of the least-squares fit line for all models under 2 Å $RMSD_{C\alpha}$ from the native complex. For homomeric complexes, 21 of the 43 complexes examined converged to the same state for all models (not necessarily at zero $RMSD_{C\alpha}$), but none of the 87 hetero-dimeric complexes did so, which demonstrated the narrowness of the funnel in homomeric complexes. For 16 of the remaining homomeric complexes and 60 hetero-dimeric complexes, where a binding funnel was recovered, the average values were calculated. As homomers generally have extensive interfaces owing to multivalent interactions, I needed to normalize the values. Dividing by the number of subunits would not account for the fact that each subunit in a homomer has more interfaces than a heterodimer. The number of interfaces is difficult to define as different homomers have different extents of interactions with non-neighboring interfaces. Instead, I normalized the slopes by dividing them by the lowest interface score observed for the complex and compared the normalized values. The radius of the funnel is defined as the difference between the models with the largest and the smallest interface $RMSD_{C\alpha}$ from the native

structure. For the complexes in which all models converged to the same $RMSD_{C\alpha}$, the funnel radius is zero. While calculating average funnel radius, I excluded complexes for which funnels could not be recovered, but included complexes with a zero funnel radius.

4.5.11. Bootstrapping

As SymDock and SymDock2 rely on random moves to dock homomers, the final output model of each trajectory is different. To produce more information about the underlying distribution of each success metric, I resample with replacement from the available model set. Bootstrapping also allows me to compare results of runs where different number of models were generated. For example, when using conformational selection from an ensemble of 155 conformations, I generated 77,500 models, but using induced fit refinement, I generated only 5,000 models. Resampling allows me to simulate selection of a desired number of models, which in this case was 2,500 models. For the various success metrics, I reported medians, averages, and standard deviations across 1,000 re-sampling attempts.

4.5.12. Success evaluation criteria

To evaluate the success of the docking simulations on the symmetric benchmark targets, I used two kinds of metrics: a near-native model count in the top-scoring models (*N#*) and fold-enrichment of near-native models in the low-scoring set ($E_{N\%}$). For example, *N5*, *N50* and *N500* are the number of near-native models in the 5, 50 and 500 top-scoring models, respectively. Fold-enrichment in the N% top-scoring models is defined as:

$$E_{N\%} = \frac{\frac{\text{\# near-native in top } N\%}{\text{\# models in top } N\%}}{\frac{\text{\# near-native}}{\text{\# models}}}$$

The bootstrapped averages for the success metrics are denoted by $\langle \cdot \rangle$.

For coarse-grained models, near-native is refined as $RMSD_{C\alpha} \leq 5$ Å. For all-atom models, near-native is defined as acceptable, medium-quality, or high-quality as per the CAPRI criteria listed below, which are based on the ligand $RMSD_{bb}$, interface $RMSD_{bb}$, and fraction of native contacts recovered.[178]

After the full protocol, if $\langle N5 \rangle \geq 3$, *i.e.* if, on average, at least 3 of the 5 top-scoring models are near-native, the complex is said to be successfully docked. This criterion was relaxed to $\langle N10 \rangle \geq 1$ when comparing to other methods to allow a fair comparison.

Chapter 5

Discussion

5.1. My contributions

The overarching goal of biomolecular complex structure modeling is to produce a structural model of the interactome. For high-throughput interaction analysis, it is essential to integrate information from experimental structure determination, bioinformatics analysis, and biophysical modeling. Computational docking methods play a key role in this pipeline by providing scalability. Two of the principal challenges that limit the accuracy of computational docking are modeling binding-induced change in the protein conformation and modeling multi-body interactions. In this thesis, I have made progress towards addressing these issues. I have improved both the sampling and the scoring aspects of the core docking protocols in the Rosetta Macromolecular Modeling Suite.[73]

In Chapter 2, I developed RosettaDock 4.0, a heterodimer docking protocol that specifically addresses the problem of flexible docking. The protocol efficiently simulates

conformational selection by sampling large, diverse backbone ensembles for both the partners while docking in a coarse-grained representation. A novel six-dimensional score function called Motif Dock Score (MDS) discriminates near-native interfaces and enriches the number of near-bound models in the coarse-grained phase. Using Rosetta's state-of-the-art all-atom score function, REF2015,[77,78] the protocol refines the all-atom models to obtain the final output structures. With ~50% accuracy in broad local docking of complexes where the interface deviated by as much as 2.2 Å, I demonstrated that the protocol can be used to predict 32% more flexible complexes than RosettaDock 3.2[104] and that it compares favorably to other leading methods. Furthermore, I showed that most current failures are a result of inadequacies in ensemble generation. As conformation generation methods improve, the docking accuracy of RosettaDock 4.0 will also improve.

In Chapter 3, I described the performance of Rosetta's docking protocols in the community-wide blind prediction experiment, Critical Assessment of PRedicted Interactions (CAPRI). My work on the targets in rounds 37–45 demonstrated the strengths and weaknesses of a variety of protocols used for docking protein homomers, heterodimers and oligosaccharides. The two glaring shortcomings that I observed were a) the set of moves in GlycanDock were too abrupt and led to large rejection rates, and b) the more tightly packed a symmetric homomer was, the more SymDock would expand the complex. While the move set in the former is being optimized presently by my colleagues, I was able to address the latter problem.

Based on the lessons of Chapter 3, in Chapter 4 I developed SymDock2, a symmetric homomer docking protocol. In this protocol, I utilized MDS to increase the near-native model counts after the coarse-grained phase of docking. Then, I explored the characteristics of the binding energy funnels for symmetric homomers and compared them to that of heterodimers. Based on my findings, I developed an induced-fit simulation in the all-atom phase. Not only does this protocol lead to six-times as many docking successes as the original SymDock,[67] its global docking performance on cyclic complexes (the most widely observed class of proteins) is significantly higher than all other methods.

In this chapter, I will discuss a) my attempts to improve heterodimer docking that did not bear fruitful results, b) the performance of SymDock2 on CAPRI target complexes had it been developed then, c) preliminary efforts and directions to utilize evolutionary data to counter the biggest bottleneck in flexible protein docking, *viz.* ensemble generation, and d) preliminary efforts and proposal to improve the modeling on dihedral symmetric complexes.

5.2. Induced fit on heterodimers

The conformational selection approach that I developed in Chapter 2 increased the limit of flexibility that we can consistently model. However, I also wanted to model the local rearrangements that take place after the formation of an encounter complex to improve model quality. Although we already repack side chains in the presence of the partner, I wanted to predict backbone rearrangements as well. The presence of a partner narrows the backbone conformational search by excluding large regions of accessible space.

Previously, inducing a better fit at the interface had been attempted by minimizing the energy of the complex along the backbone torsions of the interface residues with modest success.[61,116] I tried an alternative set of coordinates to minimize along, the Cartesian coordinates of interface residue atoms. Although much slower than torsion minimization, Cartesian minimization has previously been used to push monomer models closer to native structure and increase near-native model discrimination.[185] For the residues within 5 Å of the partner, I allowed both the side-chain and the backbone atoms to move. For residues between 5 and 10 Å, I allowed only the backbone to move. To interlace minimization moves with side-chain repacking, I used a customized relax protocol. As the atoms move independent of each other, Cartesian minimization tends to disrupt bond lengths and bond angles. To ensure that the structures remain physically realistic, I imposed constraints on bond distances, bond angles and torsion angles. Any deviation from ideal values[186] invited score penalties in the form of the following harmonic potentials:

$$E_{\text{bond\_length}} = \frac{1}{2}\sum_{i=1}^{l} k_{\text{length},i}\left(d_i - d_{i,0}\right)^2$$

$$E_{\text{bond\_angle}} = \frac{1}{2}\sum_{i=1}^{a} k_{\text{angle},i}\left(\theta_i - \theta_{i,0}\right)^2$$

$$E_{\text{bond\_torsion}} = \frac{1}{2}\sum_{i=1}^{t} k_{\text{torsion},i}\left[f_{\text{wrap}}\left(\phi_i - \phi_{i,0}, \pi\right)\right]^2$$

Here, d_i, θ_i, and ϕ_i are the i^{th} bond length, bond angle and torsion angle of the model, respectively, while $d_{i,0}$, $\theta_{i,0}$, and $\phi_{i,0}$ are the corresponding ideal values. $k_{\text{length},i}$ and $k_{\text{angle},i}$ are taken from CHARMM32,[187] and $k_{\text{torsion},i}$ is empirically derived (unpublished). f_{wrap} ensures that torsion angle difference is wrapped to the range of [0, π].

I tested this approach on 43 moderately flexible and 32 flexible complexes from the Docking Benchmark 5.0.[83] The inputs for the Cartesian minimization-based induced fit were the output models from RosettaDock 4.0. I counted the number of additional near-native models due to induced fit for medium-flexible targets (Figure 5.1) and flexible targets (Figure 5.2). While for some complexes, this approach pushed the backbones closer to the bound state, for most complexes, this was not true. The approach did not produce any additional high-quality models for any complex, and for three targets, all high-quality models were lost.

Since the binding funnel at the native interface is deeper than at false interfaces, one of the intended objectives of induced fit was to push the near-native models further into the binding funnel. I assumed that a major energy decrease would not be possible for non-native models as their energy would plateau, and hence, I sought to increase near-native discrimination by induced fit. Unfortunately, I did not observe increased discrimination. The energy scores of all models decreased relatively uniformly, which suggested that the scores were lower due to general relaxation of bond angles and bond lengths rather than specific backbone rearrangements due to the partner.

In Chapter 4, one of the principal reasons why relaxation-based induced fit worked well for symmetric homomers is that the multivalent interfaces of a subunit (leading to a deep, steep binding funnel) provide adequate guidance for inducing a fit. With just one interface, steric constraints in heterodimers are insufficient; an inter-chain clash can be resolved by a small rigid-body move rather than extensive backbone rearrangement. While this preliminary attempt did not yield encouraging results across a benchmark of protein complexes, it did improve modeling for a few of them. This suggests that in a wider induced-fit scheme, minimization along Cartesian coordinates could be one of the move sets considered. Another potential set of coordinates to move along is the normal modes of the encounter complex (as opposed to the monomers individually), which has previously been used to couple motions of both partners.[115] Complex-based normal mode analysis also provides the opportunity to incorporate induced fit in the coarse-grained phase itself, thus allowing alternating cycles of conformational selection and induced fit.

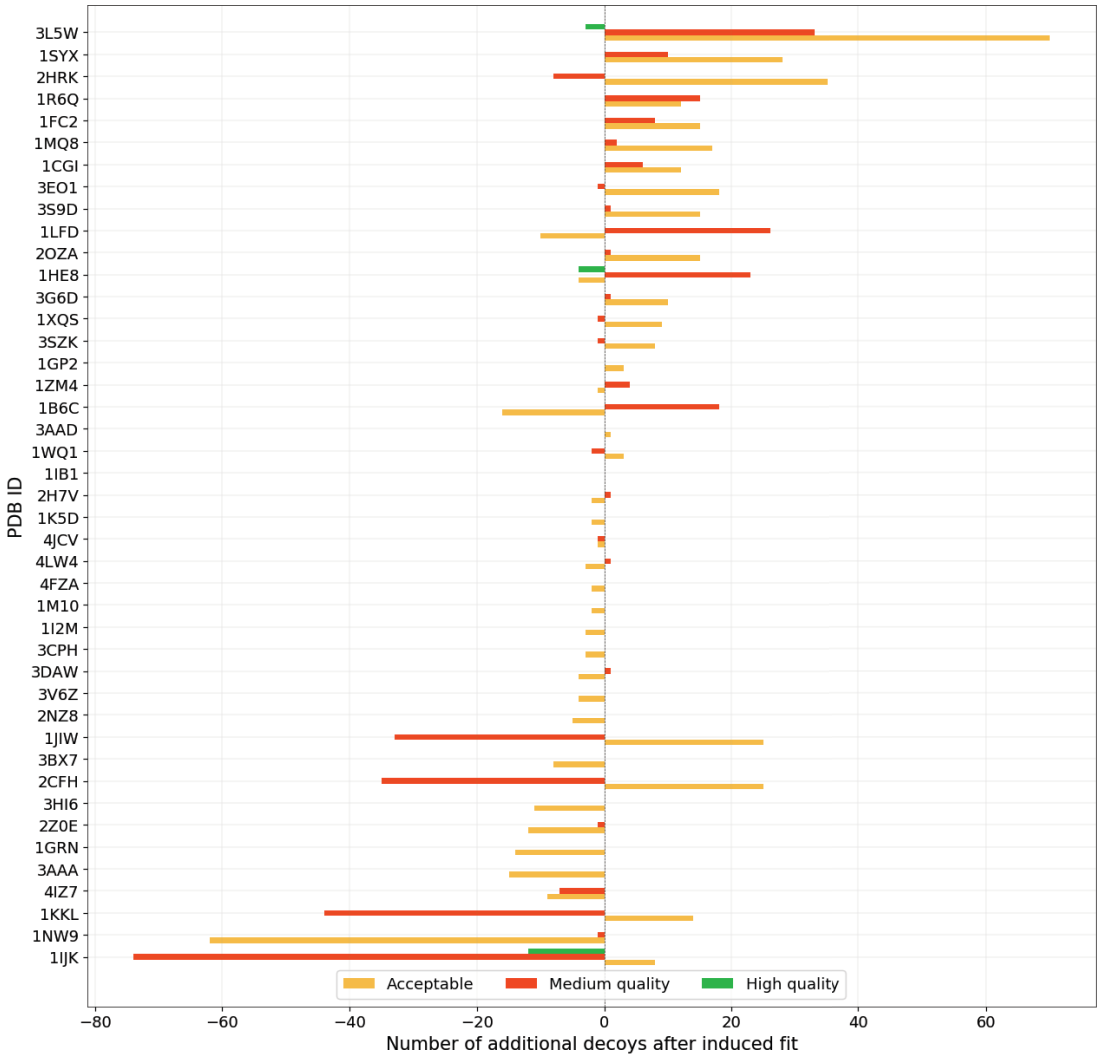


Figure 5.1: Number of additional successes after induced fit in Cartesian coordinates for medium-flexible targets. The targets are arranged in order of net gains across all categories. Unlike for symmetric homomers, induced fit does not consistently improve predictions for heterodimers. In fact, it reduces the number of high-quality models for three targets in the benchmark.

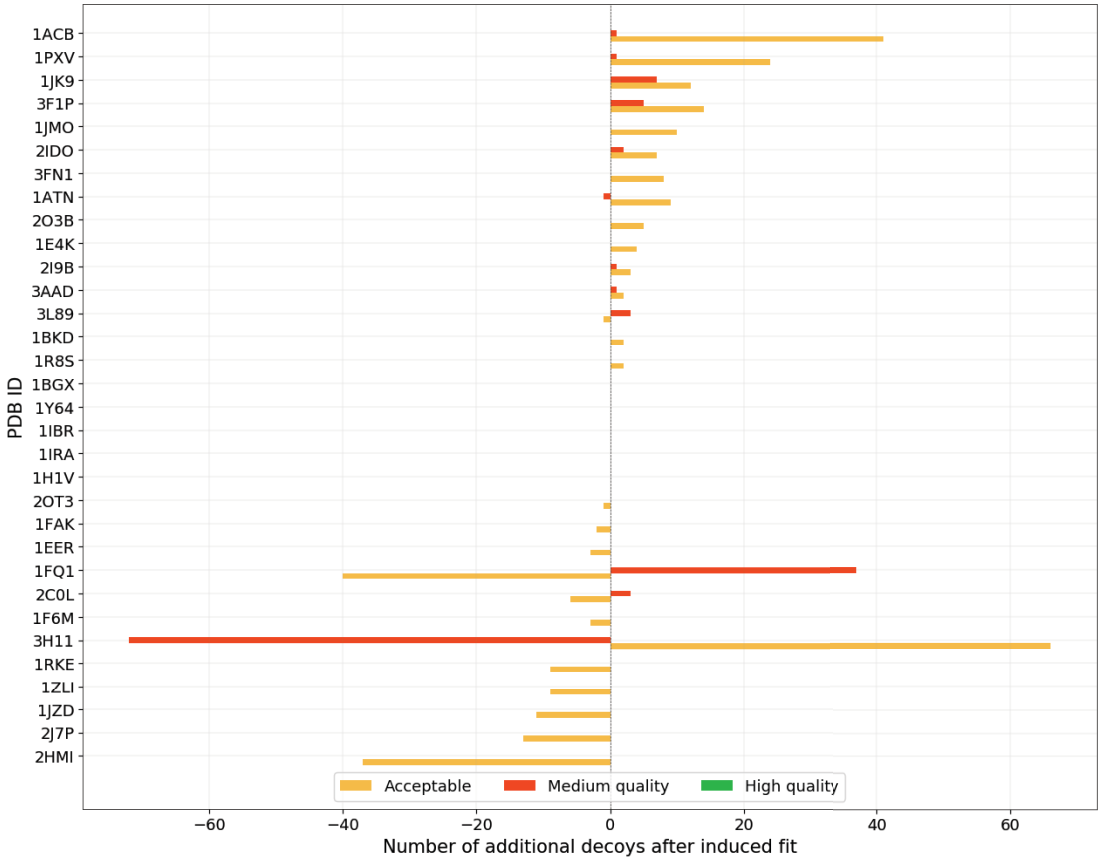


Figure 5.2: Number of additional successes after induced fit in Cartesian coordinates for highly flexible targets. The targets are arranged in order of net gains across all categories. Unlike for symmetric homomers, induced fit does not consistently improve predictions for heterodimers. For some targets, instead of causing net gains or losses, it improves acceptable models to medium-quality (*e.g.* 1FQ1), or makes medium-quality models worse (*e.g.* 3H11).

5.3. Performance of SymDock2 on CAPRI targets

In rounds 37–45, most of the successfully modeled targets were symmetric homomers. However, I had noticed a pattern of error where the Rosetta SymDock protocol would expand the overall size of the complex to relieve inter-chain clashes. This phenomenon progressively

worsened for higher order symmetries. Given the same set of inputs, I tested whether SymDock2 improved the quality of the models for three of the complexes where this error was observed in Chapter 3, *viz.* T110, T118, and T136.

For T110, the best model amongst the 10 top-scoring models had the same overall classification as our original submission, medium-quality. However, its inter-subunit distance of 23.7 Å was 2% smaller than the native (5FJL) structure's distance of 24.2 Å. This reversed the trend of the best structure from SymDock, which had a 4% larger inter-subunit distance of 25.0 Å. The cause of this shrinkage is not because the individual monomers are closer to the native; in fact, the $RMSD_{C\alpha}$ of the monomers in SymDock was 1.1 Å compared to 1.2 Å after SymDock2. The subtle backbone changes during SymDock2's flexible backbone refinement allowed for a tighter fit.

I observed a big improvement for T118, where most of SymDock2's 10 top-scoring models were high-quality (using a close homolog, 3R1M, to approximate native). The best structure from SymDock2 recovered 73% of the "native" contacts while having a sub-angstrom Lrmsd. Had I submitted this structure during CAPRI, it would have been the best structure across all groups. (Our best structure in our CAPRI submission was a medium-quality structure with 41% of the native contacts and 1.7 Å Lrmsd.) Moreover, the inter-subunit distance of the SymDock2 model was 40.3 Å compared to 43.3 Å in the SymDock model and 39.0 Å in the homolog. Thus, the SymDock2 model expands by just 3% relative to the homolog and presumably recovers additional inter-chain contacts compared to the SymDock model, which expands by 11%.

As I could not generate a feasible structure for T136 using SymDock, a direct comparison is not possible. Instead, first, I had relaxed the monomer in context of its partners and then used that monomer with SymDock. The inter-subunit distance for the best-scoring model was 62.2 Å, a 1% increase over the distance of 61.6 Å in a close homolog (2VYC). With SymDock2, I could generate models starting from the initial homology-modeled monomers with the best-scoring model having an inter-subunit distance of 63.1 Å, which was 2% more than the homolog. Thus, with flexible backbone refinement, SymDock2 resolves the problem of complex expansion in SymDock and thus, outperforms SymDock on CAPRI targets.

5.4. Future Directions

5.4.1. Flexible Protein Docking

The success of template-based and data-driven modeling methods has demonstrated that prior mined information on the structure of a complex can significantly enhance the accuracy of predictions. The data-driven score function developed in our group, MDS, exploits the frequent occurrence of residue-pair motifs at the interface to reliably identify near-native interfaces. Drawing information from the Protein Data Bank, MDS approximates the position of dozens of atoms as a single score value. This score value has more inherent information than a physical pseudoatom for approximating side chains. Just as we did for MDS in scoring, I believe that we need to find additional sources of information to inform us about protein flexibility. As I emphasized in Chapter 2, the area with the most potential for gains is

generating better ensembles. Given near-bound conformations, RosettaDock 4.0 can identify and dock them correctly.

5.4.1.1. Using inter-chain sequence co-evolution data

One data-driven method that has been successfully employed to predict structures of rigid bacterial complexes is docking guided by co-evolved residue pairs at the interface.[30,31] The conceptual underpinning for this idea is that selection pressure at protein–protein interface promotes simultaneous mutations across the interface that conserve the overall structure of the complex. Conversely, if in several homologs of target proteins *A* and *B,* positions A_i and B_j of the respective homologs co-vary, it is likely that these residues interact. All such pairs of {*Ai*, *Bj*} can be used as constraints to guide docking. Owing to the large number of bacteria sequenced, such relationships can be consistently identified. In other cases, such as eukaryotic proteins, where sequence co-variation data is less reliable, using these constraints as one of many score terms has led to a modest increase in docking success rates.[188]

I sought ways of utilizing sequence co-evolution information to generate bound-like ensembles for flexible proteins. The workflow is as follows: 1) using the set co-evolved residues pairs as ambiguous constraints, I orient the unbound partners, which may lead to substantial steric clashes, 2) in the presence of a rigid partner, I use an ensemble generation method to move the protein backbone of the more flexible protein while imposing the constraints. (In the following example, I used prior knowledge of the complex structure to determine which protein was the flexible partner and which one the rigid, but without *a priori*

knowledge, both options need to be considered.) Since the constraints are ambiguous, *i.e.* not all of them have to be simultaneously satisfied, several different starting orientations can be used to generate said varied ensembles.

The next challenge was identifying (or developing) a suitable conformer perturbation method. The perturbation methods discussed in Chapters 2 and 4 were not delivering the large conformational changes required to relieve the clashes while satisfying the constraints. Instead, I tested a method called Hybridize to partially refold the protein starting from a given backbone. This method is used for comparative modeling when the structure of a homolog is known and changes arising from insertions, deletions and mutations need to be factored in to generate a monomer model.[133] Hybridize divides the protein by secondary structural elements in the starting structure. The non-regular secondary structural segments are then folded based on fragment insertion and refinement. The presence of the rest of the protein provides context for the selection of fragments that retain the general overall shape. While moving the various segments of the protein, residue pair constraints are respected by adding a score penalty to moves where they are not satisfied.

As proof of concept, I chose residue pairs at the interface of the complex between transcription initiation factor TFIID and chaperone ASF1A to act as constraints and manually aligned the unbound monomers to satisfy these constraints (Figure 5.3A). Then, I generated 100 models of TFIID using the Hybridize method (Figure 5.3B). The backbone overlaps and clashes observed while aligning the unbound conformations were relieved by the Hybridize protocol in the generated ensemble. Compared to docking with TFIID ensembles made using

Relax, Backrub and NMA (Figure 5.3C), using with the same set of inputs for ASF1A and the Hybridize ensemble of TFIID produced better quality models with interface RMSD below 3.2 Å (Figure 5.3D). Moreover, the near-bound models were discriminated better than with the regular ensemble, leading to a docking success based on the *N5* metric.

While I demonstrated that the Hybridize protocol is a promising method to generate large, directed changes using inter-chain constraints, the constraints I used were derived from the crystal structure and not from sequence co-evolution. The ortholog-paralog problem in inter-chain co-variation necessitates analysis across a large number of closely-related sequences. Say in species X, proteins X_1 and X_2 are paralogs interacting with proteins x_1 and x_2, respectively, and in species Y, proteins Y_1 and Y_2 are paralogs interacting with proteins y_1 and y_2, respectively such that X_1, X_2, Y_1, and Y_2 are homologous, and x_1, x_2, y_1, and y_2 are homologous. If X and Y are distantly-related without the connecting sequences, it is difficult to determine whether X_1 is the ortholog of Y_1 or of Y_2. Consequently, estimating whether co-variation should be calculated for Y_1–y_1 or Y_2–y_1 is challenging.

Most flexible proteins in Docking Benchmark 5.0[83] are eukaryotic proteins, which do not have a sufficient number of related sequences for reliable identification of co-varying sites. As a result, the broad applicability of this method is currently limited. With the ever increasing number of sequenced eukaryotic genomes, this approach has high potential in the future. In the near-future, a benchmark of flexible proteins from bacterial groups where a large number of species have been sequenced needs to be compiled to refine and validate this approach.

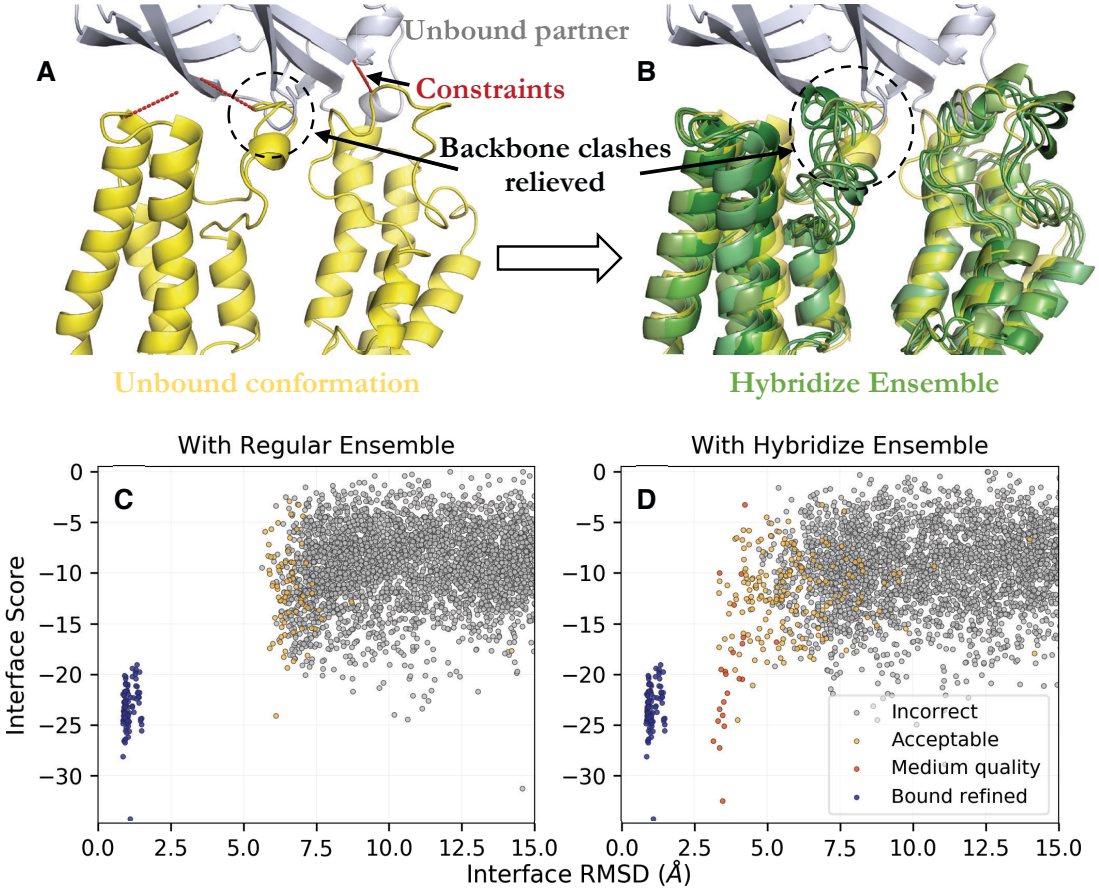


Figure 5.3: Generating conformational ensembles from inter-chain sequence co-evolution data. (A) Unbound monomers of TFIID–ASF1A complex superimposed based on constraints across the interface. Substantial steric clashes are allowed to occur. (B) Conformational ensemble of TFIID generated using the Hybridize protocol in the presence of the unbound ASF1A while respecting the imposed constraints have fewer major clashes. (C) Interface score versus interface RMSD plot shows that with ensembles of both partners generated using relax, backrub, and NMA, models fail to dock within 5 Å of the native structure. (D) Models generated using the Hybridize ensemble of TFIID docked to within 3.2 Å of the native structure.

5.4.1.2. Using monomer sequence co-evolution data

A richer source of evolutionary couplings is intra-chain sequence co-variation. Residues co-mutating across different species are likely to be in contact in the structure and important for maintaining the fold. Integrating residue-residue contact constraints with structural modeling improves protein structure prediction and has also been used to predict folds before they have been experimentally observed.[189–191] Moreover, as there is no ortholog pairing involved, reliable intra-chain sequence co-evolution data is readily obtained for a variety of species.

Residue pairs predicted to be in contact can be treated as constraints while generating ensembles, which can be especially useful in limiting the space to be sampled. Previously, Kuroda and Gray showed that current ensemble generation methods do not produce sufficiently large motions and recommended pushing the backbones further along the principal components of the overall directions of motion.[62] However, any such strategy will likely produce large motions in unintended directions and might result in corrupting the fold of the monomer. Residue pair constraints that are evolutionarily important can act as gatekeepers to maintain protein folds while allowing large motions in alterable regions.

As proof of concept, I demonstrate how evolutionary constraints derived from the GREMLIN server[189] can be used utilized to restrict loop motion in RasGAP (1WQ1). The ensemble generated using NMA, Relax, and Backrub in Chapter 2 (grey) is shown superimposed on the unbound (yellow) and the bound (green) states (Figure 5.4A). A mobile loop is highlighted with a large variety of conformations. Using C_β–C_β distance constraints on

the GREMLIN-predicted residue pairs, unnecessary motion can be restricted (Figure 5.4B). However, this ensemble did not fare any better at docking as the required type of motion was not captured by any of the ensemble generation methods.

A large-scale study of the conservation of contacts across predicted residue pairs in both unbound and bound states needs to be carried out. I expect this study to find a significantly higher conservation rate for evolutionarily-preserved contacts than randomly chosen contacts in the unbound state. If so, these constraints can form the basis of reducing false positives when using extreme perturbation strategies like high-temperature sampling or inducing large motions along normal modes or principal components of initial ensemble variation.

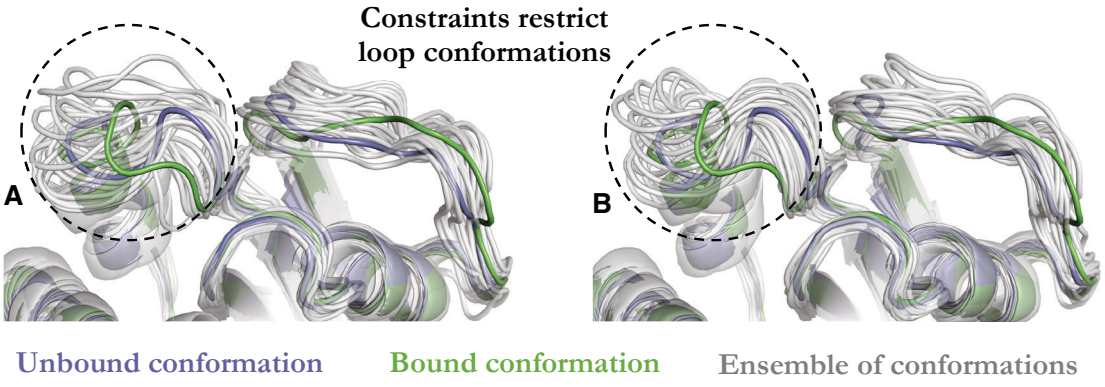


Figure 5.4: Generating conformational ensembles from intra-chain sequence co-evolution data. (A) Conformational ensemble of RasGAP generated using NMA, Relax, and Backrub from the unbound monomer superimposed on the bound state from the Ras–RasGAP complex. (B) Conformational ensemble generated using the same set of inputs and approaches, but with the additional input of predicted residue-residue contacts as C_β–C_β constraints. Adding the constraints reduces the conformational space needed to be sampled.

5.4.2. Dihedral Complex Docking

In Chapter 4, although SymDock2 greatly improved symmetric docking accuracy on dihedral homomers, the success rate of 50% was well below that of 71% for cyclic complexes. As the native state scored well for most dihedral complexes and perturbation from the native led to the recovery of binding funnels, the score function was unlikely to be the problem. As seen in Appendix Table B.2, the average lowest $RMSD_{C\alpha}$ for homology-modeled monomers of dihedral proteins was 0.9 Å, which was lower than the value of 1.4 Å of cyclic proteins, and hence, the starting monomers were not the cause of failure in most cases. Therefore, I concluded that the low accuracy was due to sampling failure of the docking protocol itself.

To test my hypothesis, I performed local docking by arranging the homology-modeled monomers as per the symmetry derived from the native complex and then randomly perturbing the subunits by 5 Å and 60°. This perturbation scrambled the subunit arrangement and interfaces, but retained the approximate ratio of the radial inter-subunit distance in each cyclic subsystem to the inter-subsystem distance. Local docking increased the median $\langle E_{1\%} \rangle$ value from 0 to 8.5 and the full protocol $\langle N5 \rangle$ value from 0.3 to 3.7. Figure 5.5 shows the docking results for two example complexes, alcohol dehydrogenase [PDBID: 1ZJZ, Sym: D2] (A and B) and KdsC phosphatase [2R8E, D4] (C and D). Both the number of near-native models and their discrimination improves.

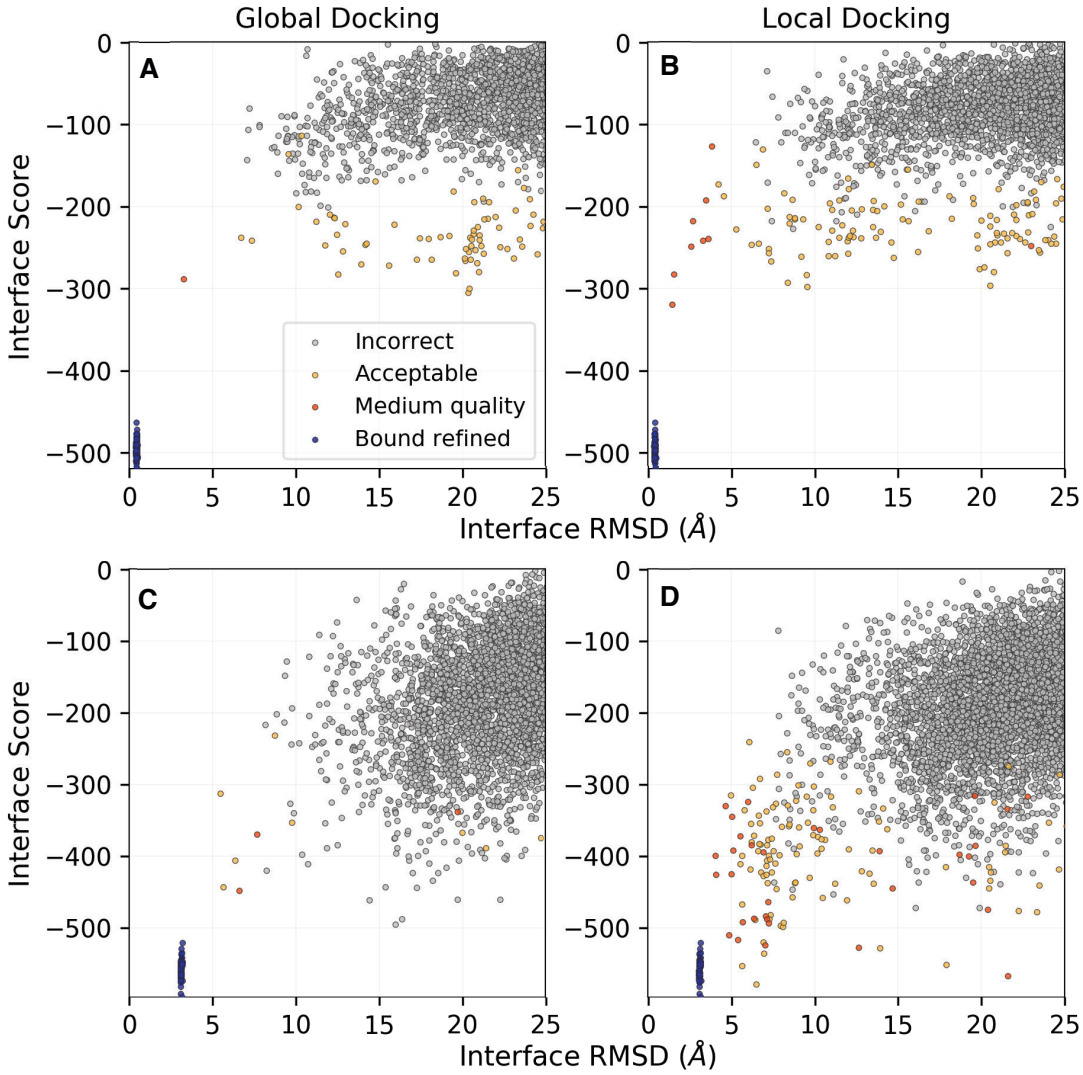


Figure 5.5: Global and local docking performance of SymDock2 on dihedral complexes. Interface score versus interface RMSD plots for (A) global docking and (B) local docking of R-specific alcohol dehydrogenase tetramer and for (C) global docking and (D) local docking of KdsC phosphatase octamer. In both the cases, starting closer to the bound state improves docking performance and near-native discrimination.

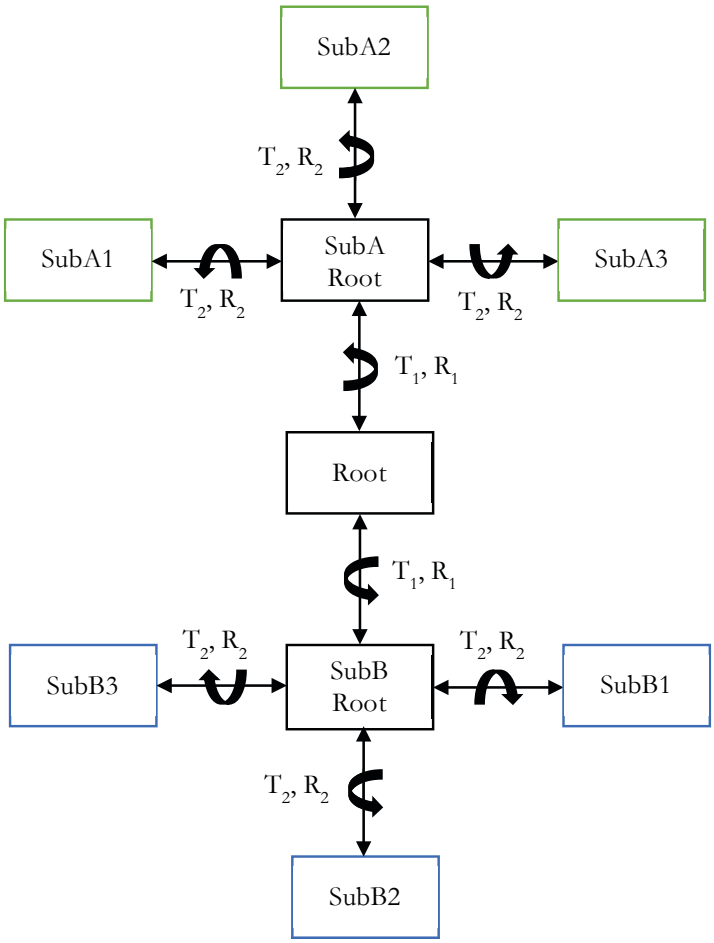


Figure 5.6: Rigid-body motion propagation order in a D3 complex. A pair of virtual atoms at the center of the complex serves as the root. The complex is divided into two sub-systems, A and B, each having its own pair of virtual root atoms. All translations (T_1) and rotations (R_1) between the *Root* and the *SubA Root* are replicated between the *Root* and the *SubB Root* in a symmetrical fashion. These motions determine the inter-sub-system distance and orientation. Each subunit has a pair of virtual atoms at its center, *e.g. SubA1*. All translations (T_2) and rotations (R_2) between *SubA* Root and *SubA1* are symmetrically replicated between the subunit virtual atoms and the respective sub-system root atoms. These motions determine radial distance (functionally equivalent to inter-subunit distance) and orientation in a sub-system.

In Rosetta, rigid-body motions between subunits are governed by a 'fold tree', which is a directed acyclic graph that determines the order of propagation of changes due to a move. A

simplified version of a D3 fold tree is presented in Figure 5.6. Translations and rotations along two sets of 'virtual atom' pairs determine inter-subunit and inter-subsystem distances and relative orientations. Presently, the same types of moves are used for both sets of motions. As I was able to obtain better a docking performance by embedding approximations of the inter-subunit and inter-subsystem distances, I see an immediate opportunity for decoupling these two move sets and optimizing their magnitudes. Perhaps a benchmark focused on dihedral homomers will be required to optimize and test the new move sets. I believe it is possible to attain a docking accuracy similar to that of cyclic complexes for dihedral complexes as well.

5.5. Conclusion

Given our current knowledge and methods, atomically-accurate interactome modeling is still a distant dream. My efforts to address the most pressing concerns in biophysical modeling of protein complexes are a key step towards realizing this dream. The scale of protein motion that needs to be captured for conformational changes of more than 2.5 Å is akin to a folding problem. This inseparability of the folding and docking is also evident for symmetric homomers with intertwined interfaces, which formed a portion of our benchmark. I believe that no single approach or source of information is enough to manage the complexity of the space that needs to be sampled. The success of future methods will rely on the ability to simultaneously leverage the strengths of various approaches such as the sampling capability of Monte Carlo simulations, the realistic kinetics of molecular dynamics simulations, and the emergent-feature recognition of unsupervised learning

Appendix A

Modeling of flexible heteromeric complexes

Table A.1: Performance of MDS vs. centroid mode for nine targets. 10,000 decoys were generated by each protocol for each target. Bootstrapped *N5*, *N100*, and *N1000* values (plus standard deviations) are listed for each target, along with average values for each metric. Cases where bootstrapping showed ≥ 50% chance of success (⟨*N5*⟩ ≥ 3, ⟨N*100*⟩ ≥ 30, or ⟨N*1000*⟩ ≥ 150) are shown in bold, and the total number of expected successes are summarized for each metric.

	Centroid			MDS		
Target	⟨*N5*⟩	⟨*N100*⟩	⟨*N1000*⟩	⟨*N5*⟩	⟨*N100*⟩	⟨*N1000*⟩
1EFN	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.6 ± 0.8	**50.0 ± 5.3**	**374 ± 16**
1GLA	0.0 ± 0.0	0.0 ± 0.0	22.1 ± 4.5	0.0 ± 0.0	5.9 ± 2.5	124 ± 10
1LFD	0.0 ± 0.0	0.0 ± 0.0	22.5 ± 4.8	**5.0 ± 0.0**	**94.9 ± 2.2**	**750 ± 18**
2A1A	0.0 ± 0.0	4.6 ± 3.5	**234 ± 14**	**2.7 ± 1.3**	**56.2 ± 5.1**	**324 ± 14**
2CFH	0.9 ± 1.2	**85.8 ± 3.5**	**673 ± 15**	**5.0 ± 0.1**	**97.9 ± 1.5**	**813 ± 13**
2FJU	1.5 ± 1.2	**83.7 ± 3.8**	**649 ± 17**	0.5 ± 0.7	**56.8 ± 5.1**	**394 ± 16**
2OT3	0.0 ± 0.0	1.1 ± 1.1	46.8 ± 6.8	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
3AAA	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.1	0.0 ± 0.0	1.8 ± 1.3	7.4 ± 2.8
3F1P	0.7 ± 0.8	13.7 ± 3.7	**337 ± 15**	**5.0 ± 0.2**	**37.8 ± 5.4**	**351 ± 16**
Average	0.3	21	220	2.1	45	349
Expected Successes	0.4	2.0	4.0	3.6	5.9	6.0

Table A.2: Performance of MDS vs. MDS without homologs for nine targets. 10,000 decoys were generated by each protocol for each target. Bootstrapped *N5*, *N100*, and *N1000* values (plus standard deviations) are listed for each target, along with average values for each metric. Cases where bootstrapping showed ≥ 50% chance of success (⟨*N5*⟩ ≥ 3, ⟨*N100*⟩ ≥ 30, or ⟨*N1000*⟩ ≥ 150) are shown in bold.

	MDS			MDS without Homologs		
Target	⟨*N5*⟩	⟨*N100*⟩	⟨*N1000*⟩	⟨*N5*⟩	⟨*N100*⟩	⟨*N1000*⟩
1EFN	0.6 ± 0.8	**50.0 ± 5.3**	**374 ± 16**	1.7 ±1.12	**57.5 ± 5.5**	**387 ± 16**
1GLA	0.0 ± 0.0	5.9 ± 2.5	124 ± 10	0.0 ± 0.0	13.0 ± 3.9	139 ± 10
1LFD	**5.0 ± 0.0**	**94.9 ± 2.2**	**750 ± 18**	**4.0 ± 1.0**	**93.8 ± 2.4**	**741 ± 17**
2A1A	**2.7 ± 1.3**	**56.2 ± 5.1**	**324 ± 14**	**3.5 ± 1.1**	**60.6 ± 5.1**	**311 ± 14**
2CFH	**5.0 ± 0.1**	**97.9 ± 1.5**	**813 ± 13**	**5.0 ± 0.0**	**100.0 ± 0.0**	**852 ± 12**
2FJU	0.5 ± 0.7	**56.8 ± 5.1**	**394 ± 16**	2.1 ± 1.2	**64.8 ± 4.9**	**422 ± 16**
2OT3	0.0 ± 0.0	0.0 ± 0.0	0 ± 0	0.0 ± 0.0	0.9 ± 0.9	5 ± 2
3AAA	0.0 ± 0.0	1.8 ± 1.3	7 ± 2	0.0 ± 0.0	2.0 ± 1.5	15 ± 3
3F1P	**5.0 ± 0.2**	**37.8 ± 5.4**	**351 ± 16**	**4.9 ± 0.3**	**53.6 ± 5.3**	**386 ± 16**
Average	2.1	45	349	2.4	49	362

Table A.3: Performance of RosettaDock 3.2 vs. RosettaDock 4.0 across an 88-target benchmark set. 5,000 decoys were generated by each protocol for each target. Bootstrapped *N5* values (plus standard deviations), both after the low-resolution phase and after the full protocol, are listed for each target. Bootstrapped enrichment values (within lowest-scoring 1% and lowest-scoring 10% of decoys) are also shown. Cases where bootstrapping gave ≥ 50% chance of success are shown in bold; success is defined as $\langle N5 \rangle \geq 3$ for the *N5* metrics, and $\langle N50 \rangle \geq 15$ and $\langle N500 \rangle \geq 75$ for the $\langle E_{1\%} \rangle$ and $\langle E_{10\%} \rangle$ metrics, respectively. For difficult targets, results from the doped RosettaDock 4.0 protocol are also shown.

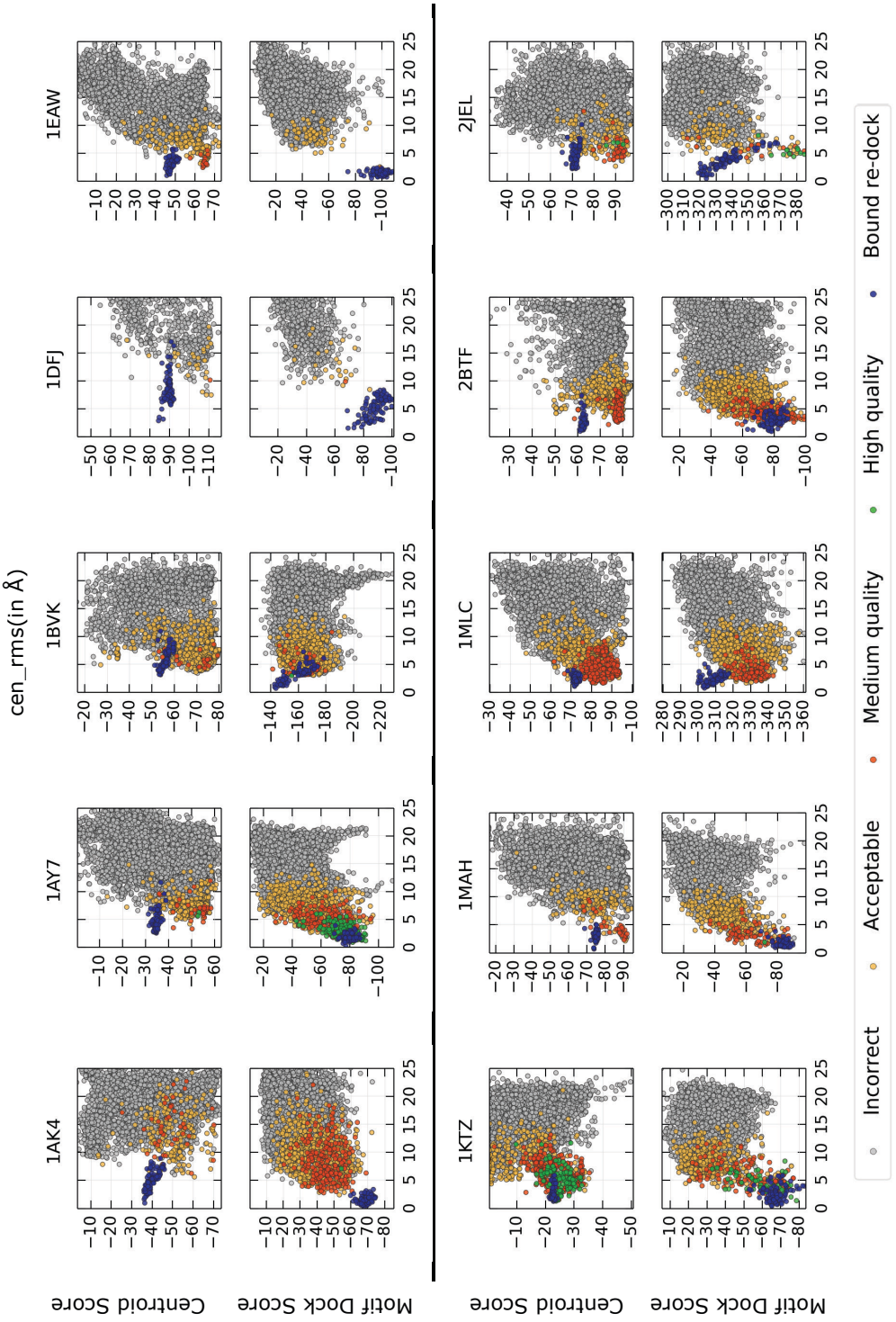
		RosettaDock 3.2				RosettaDock 4.0			
PDB	Difficulty	Low-Res $\langle N5 \rangle$	Hi-Res $\langle N5 \rangle$	$\langle E_{1\%} \rangle$	$\langle E_{10\%} \rangle$	Low-Res $\langle N5 \rangle$	Hi-Res $\langle N5 \rangle$	$\langle E_{1\%} \rangle$	$\langle E_{10\%} \rangle$
1AK4	Rigid-Body	0.0 ± 0.0	0.0 ± 0.1	0.0 ± 0.0	0.0 ± 0.9	0.0 ± 0.0	**4.0 ± 1.0**	1.2 ± 0.1	**3.8 ± 0.4**
1AY7	Rigid-Body	0.0 ± 0.0	**2.6 ± 1.2**	0.0 ± 0.0	0.0 ± 0.0	0.7 ± 0.9	**5.0 ± 0.0**	**5.1 ± 0.1**	**5.5 ± 0.2**
1BVK	Rigid-Body	0.0 ± 0.0	**4.4 ± 0.9**	0.6 ± 0.1	1.3 ± 0.2	0.0 ± 0.0	**4.8 ± 0.5**	0.0 ± 0.0	0.8 ± 0.2
1DFJ	Rigid-Body	0.0 ± 0.0	0.2 ± 0.4	N/A	N/A	0.0 ± 0.0	0.0 ± 0.0	N/A	N/A
1EAW	Rigid-Body	0.0 ± 0.0	1.0 ± 0.9	4.0 ± 0.4	3.3 ± 0.7	1.4 ± 1.1	0.0 ± 0.0	19.4 ± 1.4	4.9 ± 2.1
1KTZ	Rigid-Body	0.0 ± 0.0	**4.9 ± 0.3**	0.1 ± 0.0	0.9 ± 0.3	**4.4 ± 0.8**	**4.6 ± 0.7**	**23.7 ± 0.2**	**6.4 ± 0.6**
1MAH	Rigid-Body	0.0 ± 0.0	2.1 ± 1.2	0.8 ± 0.2	1.6 ± 1.0	2.3 ± 1.2	**4.8 ± 0.4**	**14.5 ± 0.1**	**7.3 ± 0.4**
1MLC	Rigid-Body	0.0 ± 0.0	0.2 ± 0.5	0.0 ± 0.0	**2.1 ± 0.2**	0.9 ± 0.9	0.1 ± 0.4	1.7 ± 0.1	**3.6 ± 0.2**
2BTF	Rigid-Body	0.0 ± 0.0	**5.0 ± 0.0**	0.0 ± 0.0	0.5 ± 0.2	**5.0 ± 0.0**	**5.0 ± 0.0**	**9.0 ± 0.1**	**5.1 ± 0.3**
2JEL	Rigid-Body	0.0 ± 0.0	**3.9 ± 1.2**	0.1 ± 0.1	0.6 ± 0.5	**4.4 ± 0.8**	**4.4 ± 0.9**	**50.8 ± 0.9**	8.5 ± 1.3
2PCC	Rigid-Body	0.0 ± 0.0	**3.2 ± 1.3**	0.0 ± 0.0	0.0 ± 0.3	0.0 ± 0.0	**3.0 ± 1.3**	0.0 ± 0.0	0.0 ± 0.0
2SIC	Rigid-Body	0.0 ± 0.0	**2.7 ± 1.2**	0.0 ± 0.0	1.5 ± 0.8	**5.0 ± 0.0**	**5.0 ± 0.0**	**30.8 ± 0.1**	**6.0 ± 0.6**
2SNI	Rigid-Body	0.0 ± 0.0	**5.0 ± 0.0**	0.7 ± 0.3	0.6 ± 0.7	**5.0 ± 0.0**	**5.0 ± 0.0**	**15.7 ± 0.1**	**6.8 ± 0.5**
1B6C	Medium	0.9 ± 0.9	**5.0 ± 0.0**	**2.6 ± 0.1**	**4.1 ± 0.3**	**5.0 ± 0.0**	**5.0 ± 0.0**	**19.8 ± 0.1**	**9.4 ± 0.6**
1CGI	Medium	0.0 ± 0.0	0.2 ± 0.4	0.0 ± 0.0	0.9 ± 0.1	2.0 ± 1.2	2.0 ± 1.3	**5.3 ± 0.1**	**5.4 ± 0.3**
1FC2	Medium	0.0 ± 0.0	**3.7 ± 1.0**	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.2	1.3 ± 0.1	3.0 ± 0.6
1GP2	Medium	0.0 ± 0.0	0.5 ± 0.8	0.0 ± 0.0	0.2 ± 0.2	0.0 ± 0.0	1.4 ± 1.1	29.1 ± 0.7	12.1 ± 1.7
1GRN	Medium	**2.7 ± 1.3**	2.1 ± 1.2	1.1 ± 0.1	**2.2 ± 0.4**	0.0 ± 0.0	1.3 ± 1.0	3.3 ± 0.1	**5.5 ± 0.4**
1HE8	Medium	0.0 ± 0.0	**3.3 ± 1.1**	0.0 ± 0.0	**0.7 ± 0.2**	**4.6 ± 0.7**	**4.7 ± 0.6**	**8.0 ± 0.1**	**7.8 ± 0.2**
1I2M	Medium	0.0 ± 0.0	0.8 ± 0.9	0.0 ± 0.0	1.2 ± 1.4	1.0 ± 1.0	0.0 ± 0.0	39.8 ± 1.2	9.7 ± 2.0
1IB1	Medium	0.0 ± 0.0	0.0 ± 0.0	94.7 ± 7.0	18.7 ± 9.6	0.0 ± 0.0	0.0 ± 0.0	N/A	N/A

1IJK	Medium	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	**3.7 ± 1.0**	0.0 ± 0.0	1.4 ± 0.2
1JIW	Medium	0.0 ± 0.0	0.4 ± 0.6	0.0 ± 0.0	0.2 ± 0.1	0.0 ± 0.0	0.9 ± 1.2	0.0 ± 0.0	0.8 ± 0.1
1K5D	Medium	0.0 ± 0.0	0.0 ± 0.1	0.0 ± 0.0	0.9 ± 2.3	0.0 ± 0.0	0.9 ± 0.9	0.0 ± 0.0	9.7 ± 3.6
1KKL	Medium	0.0 ± 0.0	**3.8 ± 1.0**	0.0 ± 0.0	0.3 ± 0.1	0.0 ± 0.0	**3.2 ± 1.1**	0.3 ± 0.0	**1.9 ± 0.2**
1LFD	Medium	0.0 ± 0.0	1.1 ± 1.0	0.0 ± 0.0	0.1 ± 0.2	**3.9 ± 1.0**	**5.0 ± 0.0**	**4.9 ± 0.0**	**4.5 ± 0.1**
1M10	Medium	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.1	0.0 ± 0.0	6.7 ± 1.3
1MQ8	Medium	0.0 ± 0.0	**4.9 ± 0.3**	0.0 ± 0.0	0.8 ± 0.3	0.0 ± 0.0	**5.0 ± 0.0**	0.3 ± 0.0	2.4 ± 0.3
1NW9	Medium	0.0 ± 0.0	1.3 ± 1.1	0.0 ± 0.0	0.5 ± 0.2	1.3 ± 1.0	**3.6 ± 1.1**	**3.6 ± 0.1**	**4.4 ± 0.2**
1R6Q	Medium	0.0 ± 0.1	1.8 ± 1.1	0.8 ± 0.1	**1.4 ± 0.2**	2.1 ± 1.2	**2.7 ± 1.3**	**5.6 ± 0.1**	**4.5 ± 0.2**
1SYX	Medium	**3.7 ± 1.0**	**5.0 ± 0.0**	**2.5 ± 0.1**	0.6 ± 0.2	0.0 ± 0.0	**5.0 ± 0.0**	0.0 ± 0.0	**1.4 ± 0.1**
1WQ1	Medium	0.0 ± 0.0	**2.9 ± 1.2**	0.0 ± 0.0	0.2 ± 0.1	0.5 ± 0.8	1.9 ± 1.5	**15.7 ± 0.3**	**10.7 ± 0.8**
1XQS	Medium	0.0 ± 0.0	**3.9 ± 1.0**	0.1 ± 0.1	3.0 ± 0.9	1.5 ± 1.1	**4.3 ± 0.9**	14.4 ± 0.4	**12.4 ± 1.2**
1ZM4	Medium	0.0 ± 0.0	1.0 ± 1.0	0.0 ± 0.0	1.9 ± 0.6	1.3 ± 1.1	2.5 ± 1.2	12.3 ± 0.3	4.0 ± 0.6
2CFH	Medium	**5.0 ± 0.0**	**4.3 ± 0.8**	**6.5 ± 0.0**	**5.2 ± 0.2**	**4.1 ± 0.9**	**5.0 ± 0.0**	**10.2 ± 0.0**	**8.4 ± 0.2**
2H7V	Medium	0.0 ± 0.0	**2.8 ± 1.3**	0.0 ± 0.0	0.0 ± 0.0	**4.1 ± 0.9**	**2.6 ± 1.1**	**28.3 ± 0.4**	**8.2 ± 0.9**
2HRK	Medium	0.9 ± 0.9	1.0 ± 1.0	0.4 ± 0.1	0.8 ± 0.3	0.0 ± 0.0	**4.3 ± 0.9**	0.5 ± 0.0	1.3 ± 0.2
2NZ8	Medium	0.0 ± 0.0	2.3 ± 1.3	0.0 ± 0.0	0.1 ± 0.1	0.0 ± 0.0	**3.6 ± 1.1**	1.0 ± 0.1	**3.9 ± 0.3**
2OZA	Medium	0.0 ± 0.0	0.2 ± 0.5	0.2 ± 0.1	1.8 ± 0.7	**2.9 ± 1.1**	0.0 ± 0.1	**17.6 ± 0.3**	6.4 ± 0.8
2Z0E	Medium	0.0 ± 0.0	0.1 ± 0.3	0.0 ± 0.0	0.5 ± 0.5	0.0 ± 0.1	0.9 ± 0.9	6.1 ± 0.3	4.0 ± 0.9
3AAA	Medium	0.0 ± 0.0	0.1 ± 0.3	0.0 ± 0.0	0.5 ± 1.9	0.0 ± 0.0	1.5 ± 1.1	0.0 ± 0.0	2.2 ± 0.7
3AAD	Medium	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.3
3BX7	Medium	0.0 ± 0.0	0.0 ± 0.0	0.5 ± 0.0	**6.2 ± 0.3**	2.4 ± 1.2	**2.5 ± 1.3**	**9.4 ± 0.2**	**7.2 ± 0.6**
3CPH	Medium	0.0 ± 0.0	**2.8 ± 1.5**	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.4 ± 0.7	0.0 ± 0.0	1.5 ± 0.3
3DAW	Medium	0.0 ± 0.0	**4.7 ± 0.7**	0.6 ± 0.2	0.6 ± 0.6	1.8 ± 1.1	**4.5 ± 0.7**	**18.1 ± 0.3**	**8.2 ± 0.7**
3EO1	Medium	0.0 ± 0.0	**5.0 ± 0.0**	0.0 ± 0.0	1.2 ± 0.4	0.0 ± 0.0	**5.0 ± 0.2**	0.0 ± 0.0	5.1 ± 0.6
3G6D	Medium	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	1.1 ± 1.0	0.2 ± 0.0	**3.1 ± 0.3**
3HI6	Medium	0.0 ± 0.0	0.0 ± 0.0	0.3 ± 0.0	0.3 ± 0.1	0.0 ± 0.2	0.0 ± 0.2	4.6 ± 0.2	4.9 ± 0.8
3L5W	Medium	0.0 ± 0.0	0.5 ± 1.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	**5.0 ± 0.0**	0.0 ± 0.0	**2.5 ± 0.2**
3S9D	Medium	0.0 ± 0.0	**3.9 ± 1.0**	0.0 ± 0.0	0.2 ± 0.3	0.9 ± 1.0	**3.9 ± 1.1**	**12.7 ± 0.2**	**6.8 ± 0.5**
3SZK	Medium	0.0 ± 0.0	**3.3 ± 1.2**	0.0 ± 0.0	0.4 ± 0.3	1.0 ± 1.0	**3.3 ± 1.2**	4.0 ± 0.1	3.7 ± 0.4
3V6Z	Medium	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
4FZA	Medium	0.0 ± 0.0	0.4 ± 0.7	0.0 ± 0.0	2.0 ± 0.7	0.0 ± 0.0	1.0 ± 1.0	0.0 ± 0.0	4.2 ± 0.5

4IZ7	Medium	0.0 ± 0.0	0.2 ± 0.4	0.0 ± 0.0	0.1 ± 0.1	0.0 ± 0.0	0.0 ± 0.1	0.0 ± 0.0	0.6 ± 0.2
4JCV	Medium	0.0 ± 0.0	0.0 ± 0.0	3.6 ± 1.9	1.1 ± 3.4	0.0 ± 0.2	**2.5 ± 1.2**	19.4 ± 1.3	8.8 ± 2.8
4LW4	Medium	0.0 ± 0.0	**3.9 ± 1.1**	0.3 ± 0.3	0.3 ± 0.9	1.0 ± 0.9	1.5 ± 1.1	26.8 ± 0.7	8.4 ± 1.4
1ACB	Difficult	0.0 ± 0.0	**4.1 ± 1.0**	0.0 ± 0.0	0.2 ± 0.3	0.2 ± 0.5	2.4 ± 1.2	0.4 ± 0.0	1.3 ± 0.2
1ACB	Doped					0.7 ± 0.9	**4.4 ± 0.9**	3.8 ± 0.1	1.9 ± 0.3
1ATN	Difficult	0.0 ± 0.0	**4.3 ± 0.9**	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	**2.8 ± 1.4**	0.0 ± 0.0	7.5 ± 1.2
1ATN	Doped					0.0 ± 0.0	**3.2 ± 1.2**	0.0 ± 0.0	5.7 ± 1.1
1BGX	Difficult	0.0 ± 0.0	0.0 ± 0.0	N/A	N/A	0.0 ± 0.0	0.0 ± 0.0	N/A	N/A
1BGX	Doped					0.0 ± 0.0	0.0 ± 0.0	N/A	N/A
1BKD	Difficult	0.0 ± 0.0	0.0 ± 0.0	0.6 ± 0.4	2.0 ± 1.8	0.0 ± 0.0	0.0 ± 0.0	20.0 ± 1.3	8.8 ± 2.7
1BKD	Doped					0.0 ± 0.0	0.0 ± 0.0	16.2 ± 1.5	9.7 ± 3.5
1E4K	Difficult	0.1 ± 0.2	0.0 ± 0.0	2.7 ± 0.2	7.0 ± 0.7	2.1 ± 1.2	0.0 ± 0.2	17.9 ± 0.5	6.7 ± 1.0
1E4K	Doped					0.0 ± 0.0	0.0 ± 0.1	2.6 ± 0.3	3.1 ± 0.9
1EER	Difficult	0.0 ± 0.0	0.0 ± 0.0	N/A	N/A	0.0 ± 0.0	0.1 ± 0.4	N/A	N/A
1EER	Doped					0.0 ± 0.0	0.7 ± 0.9	N/A	N/A
1F6M	Difficult	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.3 ± 0.7	0.0 ± 0.0	0.6 ± 0.2
1F6M	Doped					0.6 ± 0.9	**3.8 ± 1.2**	3.2 ± 0.1	1.4 ± 0.2
1FAK	Difficult	0.0 ± 0.0	0.0 ± 0.1	54.3 ± 1.0	23.8 ± 2.1	0.0 ± 0.0	0.0 ± 0.2	0.0 ± 0.0	1.8 ± 1.8
1FAK	Doped					0.0 ± 0.0	0.0 ± 0.0	31.4 ± 2.3	8.9 ± 3.9
1FQ1	Difficult	0.0 ± 0.0	**4.8 ± 0.5**	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	**5.0 ± 0.1**	2.4 ± 0.2	2.8 ± 0.6
1FQ1	Doped					0.0 ± 0.0	**5.0 ± 0.0**	0.0 ± 0.0	3.4 ± 0.6
1H1V	Difficult	0.0 ± 0.0	0.0 ± 0.0	N/A	N/A	0.0 ± 0.0	0.0 ± 0.0	N/A	N/A
1H1V	Doped					0.0 ± 0.0	0.0 ± 0.0	N/A	N/A
1IBR	Difficult	0.0 ± 0.0	0.0 ± 0.0	N/A	N/A	0.0 ± 0.0	0.0 ± 0.0	N/A	N/A
1IBR	Doped					0.0 ± 0.0	0.0 ± 0.0	N/A	N/A
1IRA	Difficult	0.0 ± 0.0	0.0 ± 0.0	N/A	N/A	0.0 ± 0.0	0.0 ± 0.0	N/A	N/A
1IRA	Doped					0.0 ± 0.0	0.0 ± 0.0	N/A	N/A
1JK9	Difficult	0.0 ± 0.0	**5.0 ± 0.0**	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.1	**5.0 ± 0.0**	**7.0 ± 0.1**	**4.4 ± 0.4**
1JK9	Doped					0.0 ± 0.0	**5.0 ± 0.0**	**6.2 ± 0.1**	**3.7 ± 0.3**
1JMO	Difficult	0.0 ± 0.0	1.0 ± 1.0	0.0 ± 0.0	3.7 ± 0.9	0.0 ± 0.0	1.4 ± 1.0	0.1 ± 0.0	0.5 ± 0.2
1JMO	Doped					0.0 ± 0.0	**4.7 ± 0.7**	0.5 ± 0.1	1.1 ± 0.2
1JZD	Difficult	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	1.8 ± 1.3	7.5 ± 0.3	6.3 ± 0.8

1JZD	Doped					0.0 ± 0.2	1.2 ± 1.0	10.2 ± 0.3	6.5 ± 0.8
1PXV	Difficult	0.0 ± 0.0	0.0 ± 0.0	0.8 ± 0.1	5.1 ± 0.5	**3.3 ± 1.2**	**2.5 ± 1.3**	**21.9 ± 0.4**	**7.7 ± 0.9**
1PXV	Doped					**4.7 ± 0.6**	**4.6 ± 0.8**	**29.7 ± 0.3**	**8.1 ± 0.8**
1R8S	Difficult	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	33.3 ± 3.2	3.3 ± 3.2
1R8S	Doped					**4.1 ± 1.0**	2.0 ± 1.2	67.3 ± 2.6	8.6 ± 2.9
1RKE	Difficult	0.0 ± 0.0	0.0 ± 0.0	4.6 ± 0.4	3.4 ± 1.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.5 ± 0.2
1RKE	Doped					0.0 ± 0.0	0.6 ± 0.8	0.0 ± 0.0	0.3 ± 0.2
1Y64	Difficult	0.0 ± 0.0	0.0 ± 0.0	N/A	N/A	0.0 ± 0.0	0.0 ± 0.0	N/A	N/A
1Y64	Doped					0.0 ± 0.0	0.0 ± 0.0	N/A	N/A
1ZLI	Difficult	0.0 ± 0.2	**3.6 ± 1.1**	0.8 ± 0.2	1.0 ± 0.7	2.1 ± 1.2	0.7 ± 0.8	24.4 ± 0.9	8.3 ± 1.7
1ZLI	Doped					1.0 ± 1.0	1.2 ± 1.0	32.6 ± 0.8	8.2 ± 1.5
2C0L	Difficult	0.0 ± 0.0	1.3 ± 1.0	0.0 ± 0.0	0.3 ± 0.5	0.4 ± 0.7	**4.0 ± 1.0**	2.6 ± 0.1	2.7 ± 0.4
2C0L	Doped					0.0 ± 0.0	0.2 ± 0.5	0.0 ± 0.0	1.4 ± 0.5
2HMI	Difficult	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	**3.0 ± 1.3**	0.0 ± 0.0	0.8 ± 0.5
2HMI	Doped					0.0 ± 0.0	**3.3 ± 1.3**	0.0 ± 0.0	1.2 ± 0.7
2I9B	Difficult	0.0 ± 0.0	1.0 ± 0.9	0.0 ± 0.0	1.2 ± 1.4	1.3 ± 1.2	0.4 ± 0.8	**40.4 ± 0.9**	**17.9 ± 1.8**
2I9B	Doped					0.8 ± 0.9	**2.7 ± 1.2**	11.0 ± 0.4	9.2 ± 1.1
2IDO	Difficult	0.0 ± 0.0	0.0 ± 0.1	0.0 ± 0.0	0.2 ± 0.5	0.0 ± 0.0	**3.8 ± 1.1**	0.0 ± 0.0	1.4 ± 0.3
2IDO	Doped					0.0 ± 0.0	**3.9 ± 1.0**	0.0 ± 0.0	1.4 ± 0.3
2J7P	Difficult	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	2.3 ± 0.2	0.5 ± 0.3
2J7P	Doped					0.0 ± 0.0	**4.9 ± 0.3**	12.0 ± 0.3	5.5 ± 0.6
2O3B	Difficult	0.0 ± 0.0	1.1 ± 1.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	1.4 ± 1.2	0.0 ± 0.0	1.2 ± 0.5
2O3B	Doped					0.0 ± 0.0	1.8 ± 1.1	0.0 ± 0.0	0.4 ± 0.3
2OT3	Difficult	0.0 ± 0.0	0.0 ± 0.0	6.2 ± 0.8	4.4 ± 2.1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
2OT3	Doped					1.0 ± 1.0	1.4 ± 1.0	16.3 ± 1.0	3.8 ± 1.5
3AAA	Difficult	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.3
3AAA	Doped					0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.4 ± 0.4
3F1P	Difficult	0.0 ± 0.0	1.1 ± 1.0	0.0 ± 0.0	**1.9 ± 0.2**	**3.9 ± 1.0**	**5.0 ± 0.1**	**6.1 ± 0.1**	**4.2 ± 0.3**
3F1P	Doped					**4.6 ± 0.7**	**4.6 ± 0.8**	**7.1 ± 0.1**	**4.8 ± 0.3**
3FN1	Difficult	0.0 ± 0.0	1.8 ± 1.3	1.5 ± 0.2	**1.5 ± 0.7**	0.0 ± 0.0	1.6 ± 1.2	7.7 ± 0.4	7.0 ± 1.0
3FN1	Doped					0.5 ± 0.7	0.9 ± 1.0	4.9 ± 0.2	2.9 ± 0.5
3H11	Difficult	0.0 ± 0.1	**4.0 ± 0.9**	1.8 ± 0.1	0.8 ± 0.3	**5.0 ± 0.0**	**5.0 ± 0.0**	**10.6 ± 0.0**	**8.7 ± 0.3**

3H11	Doped					**4.0 ± 0.9**	**5.0 ± 0.0**	**9.0 ± 0.0**	**8.1 ± 0.2**
3L89	Difficult	0.0 ± 0.0	**5.0 ± 0.2**	0.5 ± 0.1	**1.6 ± 0.4**	0.0 ± 0.0	**4.3 ± 0.8**	16.6 ± 0.4	7.9 ± 1.0
3L89	Doped					0.9 ± 0.9	**4.2 ± 0.8**	11.8 ± 0.3	**8.1 ± 0.9**


cen_rms(in Å)
1AK4
1AY7
1BVK
1DFJ
1EAW
1KTZ
1MAH
1MLC
2BTF
2JEL
Centroid Score
Motif Dock Score
Centroid Score
Motif Dock Score
Incorrect
Acceptable
Medium quality
High quality
Bound re-dock

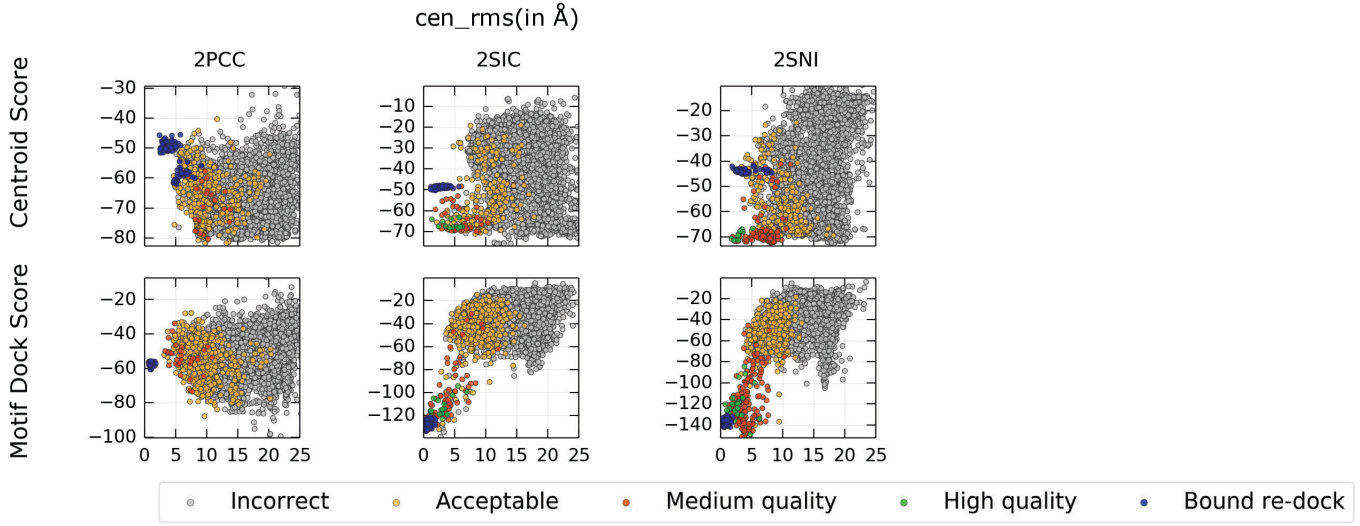
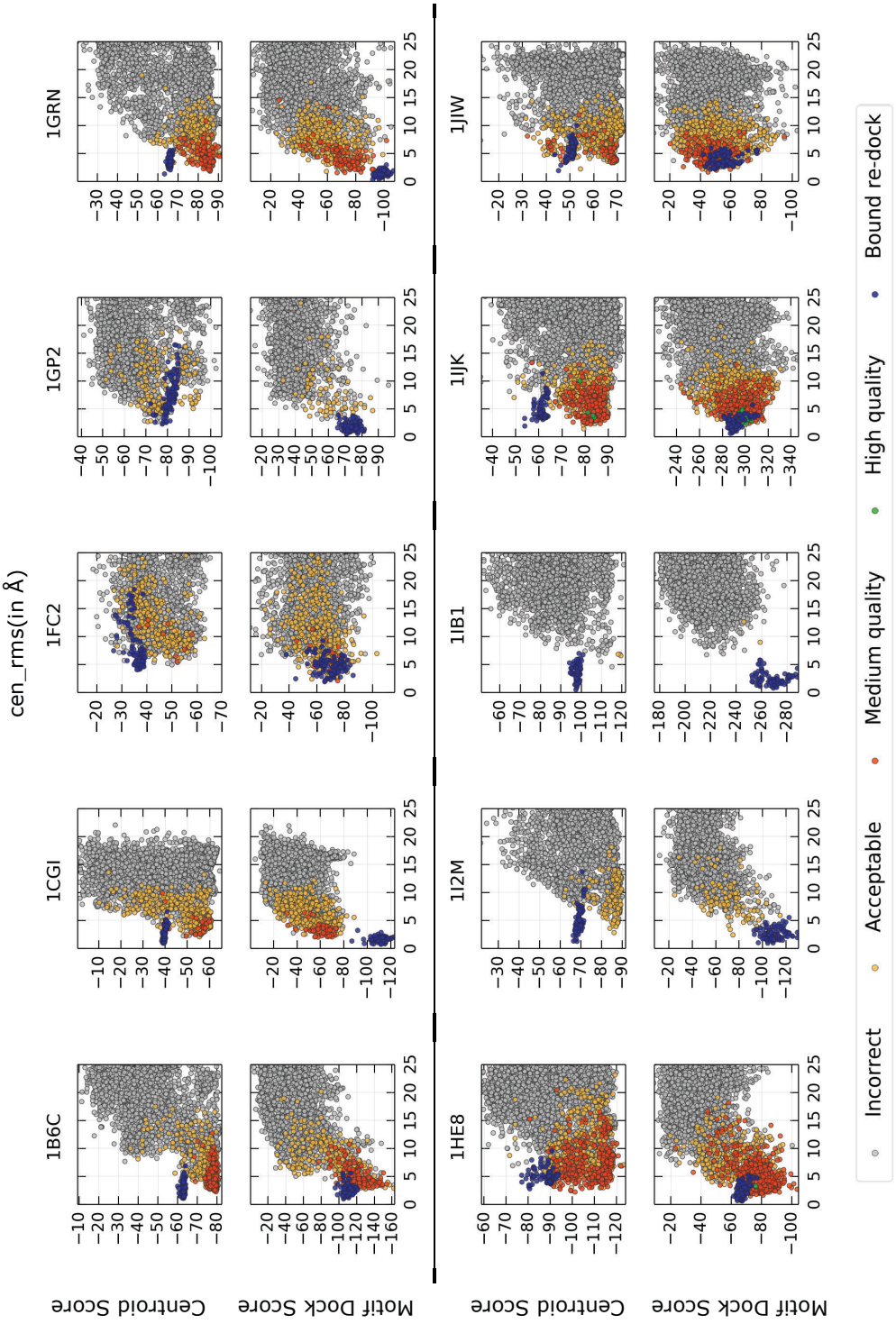
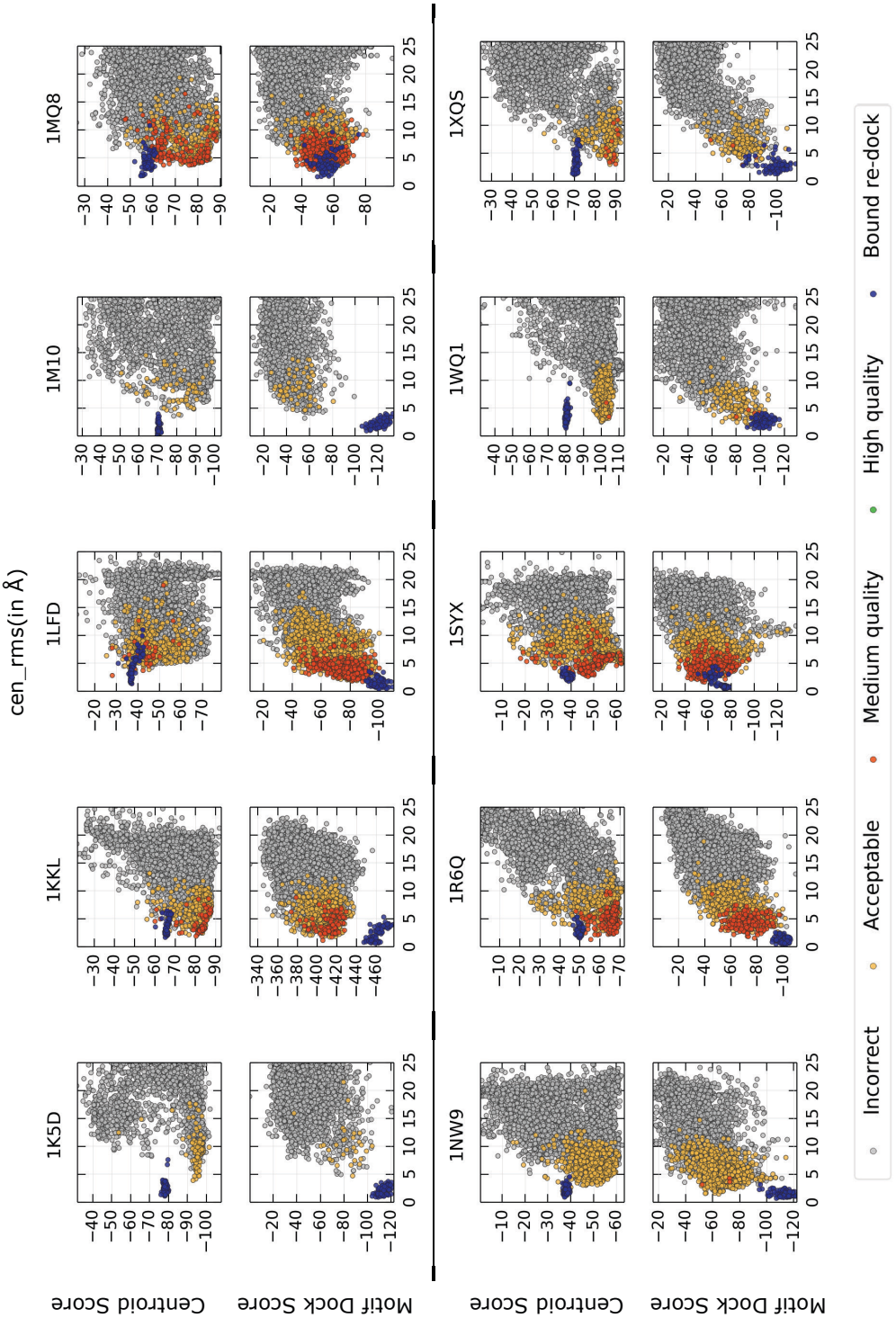


Figure A.1: Score versus RMSD plots in the low-resolution stage for centroid score versus motif dock score for rigid complexes.


cen_rms(in Å)
1B6C
1CGI
1FC2
1GP2
1GRN
1HE8
1I2M
1IB1
1IJK
1JIW
Centroid Score
Motif Dock Score
Incorrect
Acceptable
Medium quality
High quality
Bound re-dock


cen_rms(in Å)
1K5D
1KKL
1LFD
1M10
1MQ8
1NW9
1R6Q
1SYX
1WQ1
1XQS
Centroid Score
Motif Dock Score
Incorrect
Acceptable
Medium quality
High quality
Bound re-dock

cen_rms(in Å)
1ZM4
2CFH
2H7V
2HRK
2NZ8
2OZA
2Z0E
3AAA
3AAD
3BX7
Centroid Score
Motif Dock Score
Incorrect
Acceptable
Medium quality
High quality
Bound re-dock

cen_rms(in Å)
3CPH
3DAW
3EO1
3G6D
3HI6
3L5W
3S9D
3SZK
3V6Z
4FZA
Centroid Score
Motif Dock Score
Incorrect
Acceptable
Medium quality
High quality
Bound re-dock

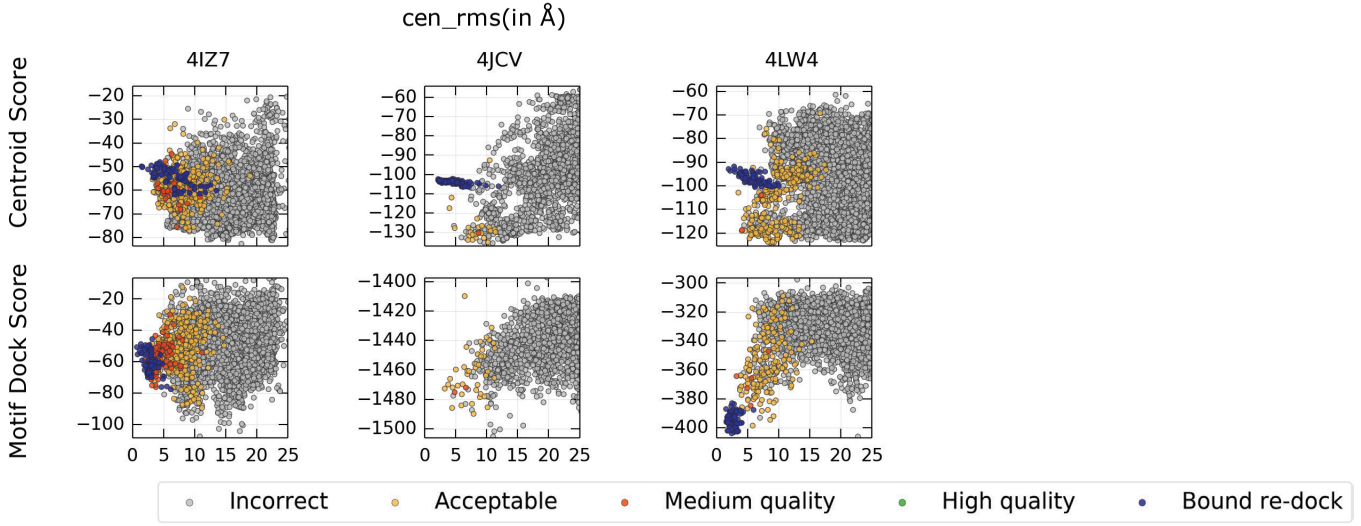
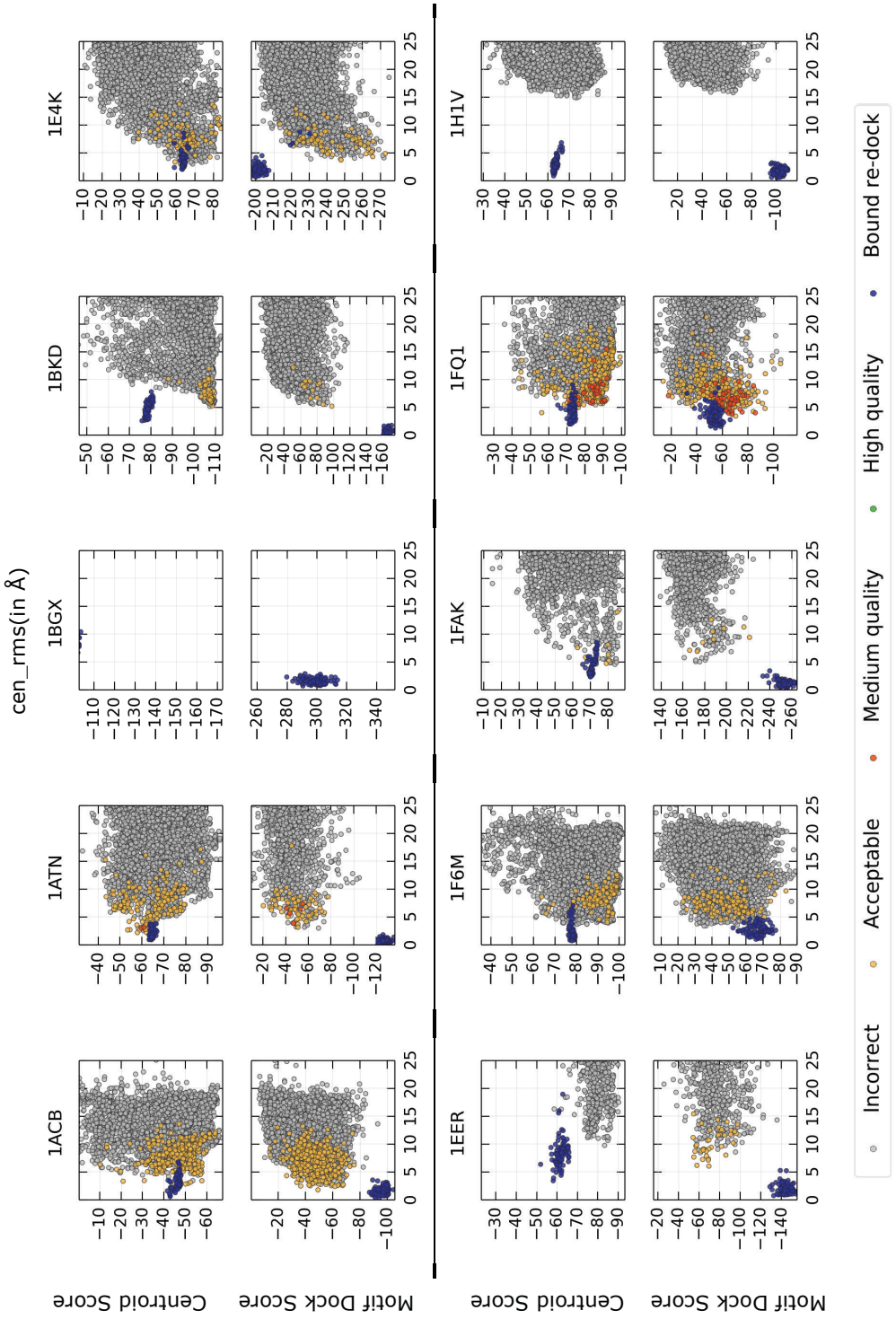
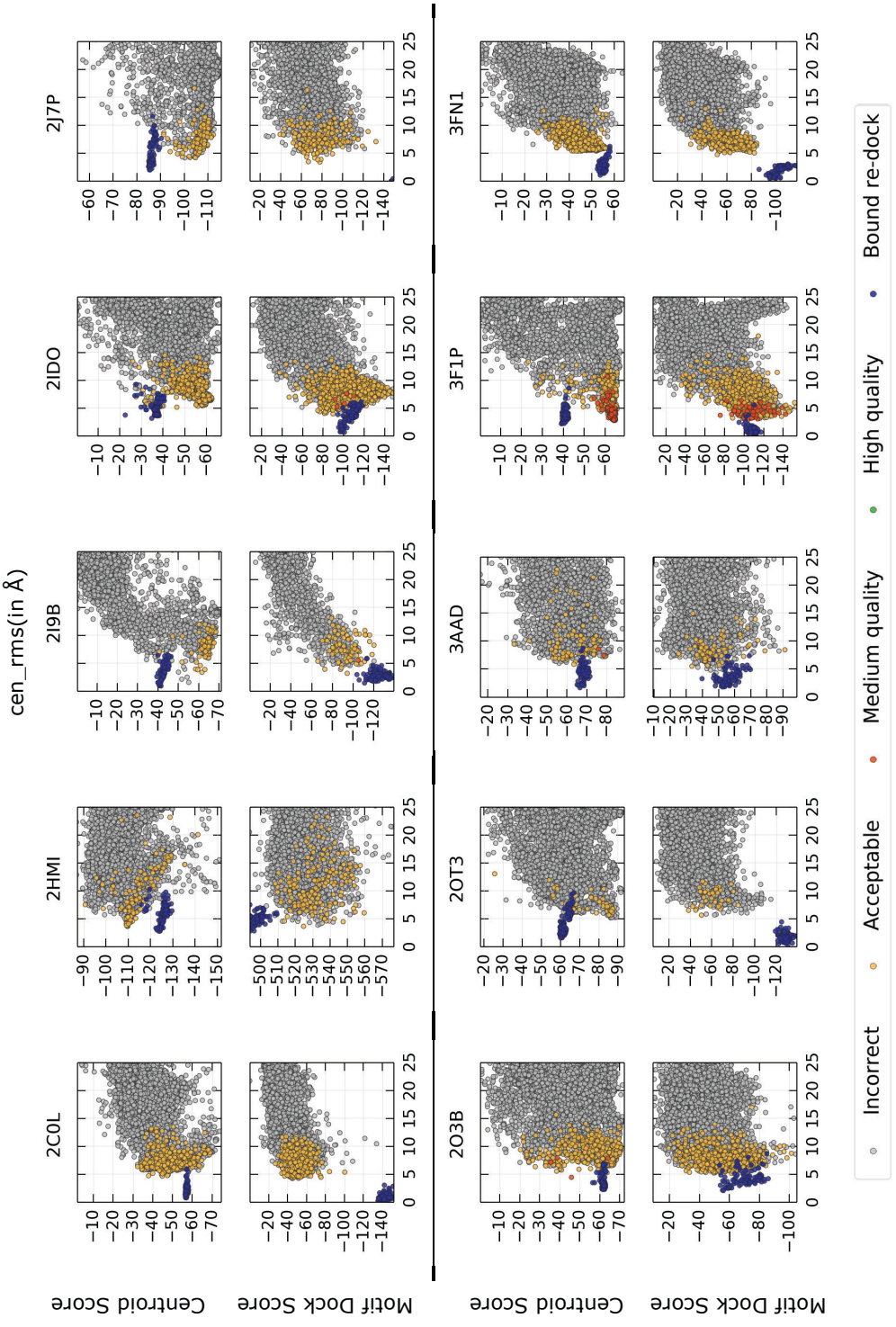


Figure A.2: Score versus RMSD plots in the low-resolution stage for centroid score versus motif dock score for medium-flexibility complexes.


cen_rms(in Å)
1ACB
1ATN
1BGX
1BKD
1E4K
1EER
1F6M
1FAK
1FQ1
1H1V
Centroid Score
Motif Dock Score
Centroid Score
Motif Dock Score
Incorrect
Acceptable
Medium quality
High quality
Bound re-dock

cen_rms(in Å)
1IBR
1IRA
1JK9
1JMO
1JZD
1PXV
1R8S
1RKE
1Y64
1ZLI
Centroid Score
Motif Dock Score
Incorrect
Acceptable
Medium quality
High quality
Bound re-dock


cen_rms(in Å)
2C0L
2HMI
2I9B
2IDO
2J7P
2O3B
2OT3
3AAD
3F1P
3FN1
Centroid Score
Motif Dock Score
Incorrect
Acceptable
Medium quality
High quality
Bound re-dock

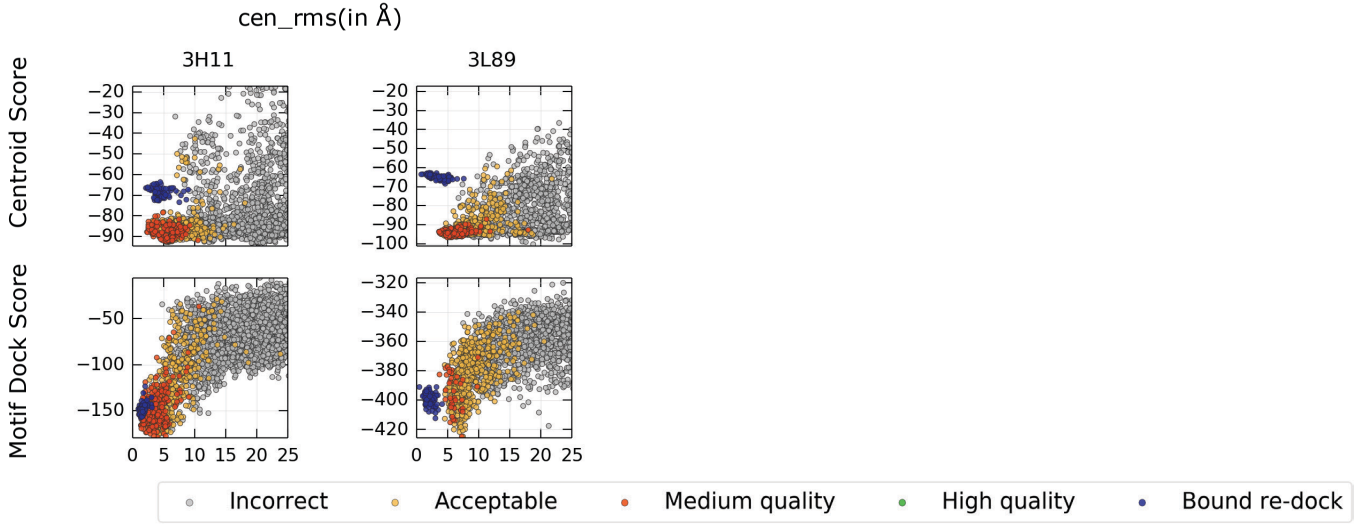
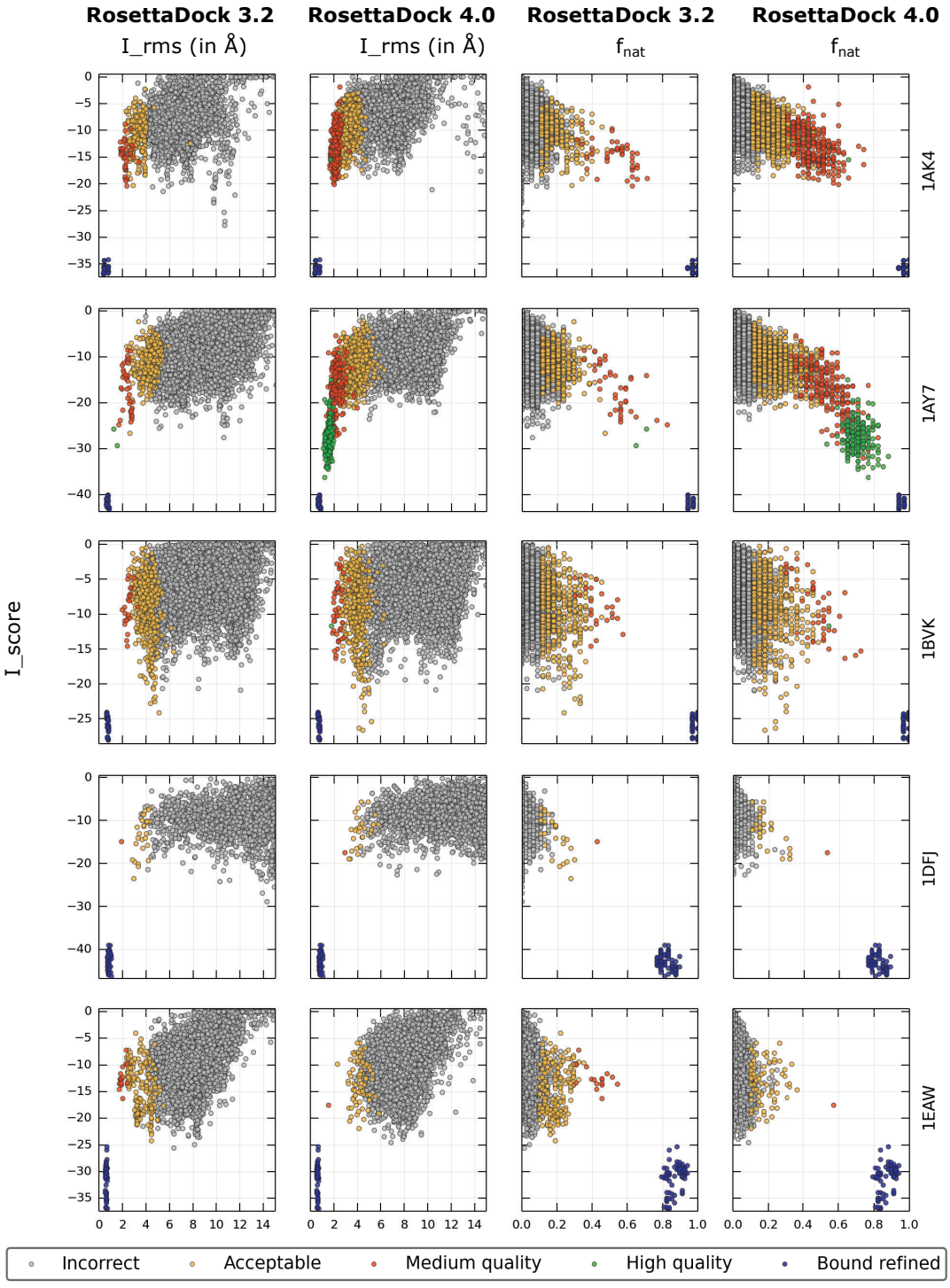
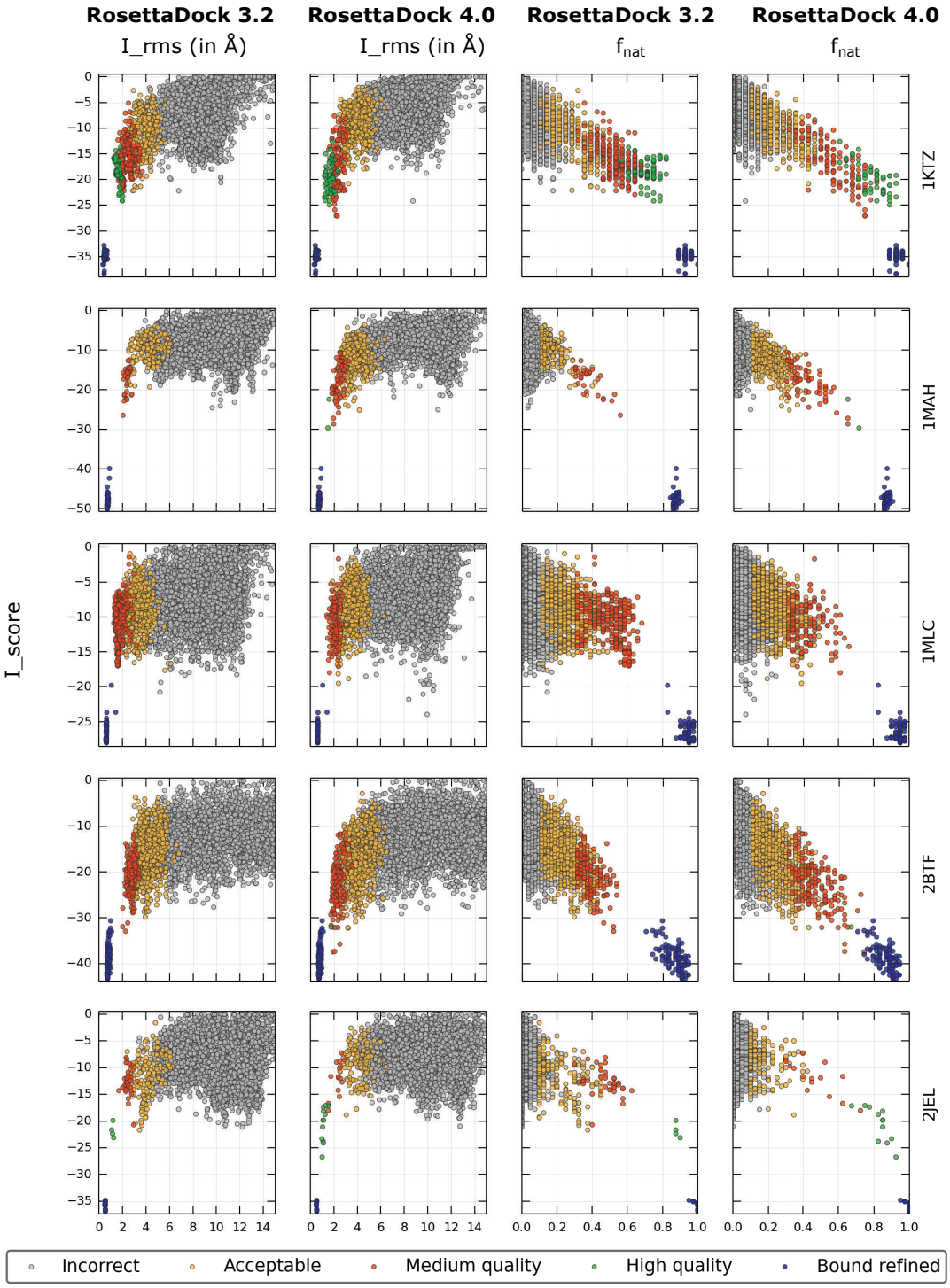


Figure A.3: Score versus RMSD plots in the low-resolution stage for centroid score versus motif dock score for flexible complexes.


RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
I_score
1AK4
1AY7
1BVK
1DFJ
1EAW
Incorrect
Acceptable
Medium quality
High quality
Bound refined


RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
I_score
1KTZ
1MAH
1MLC
2BTF
2JEL
Incorrect
Acceptable
Medium quality
High quality
Bound refined

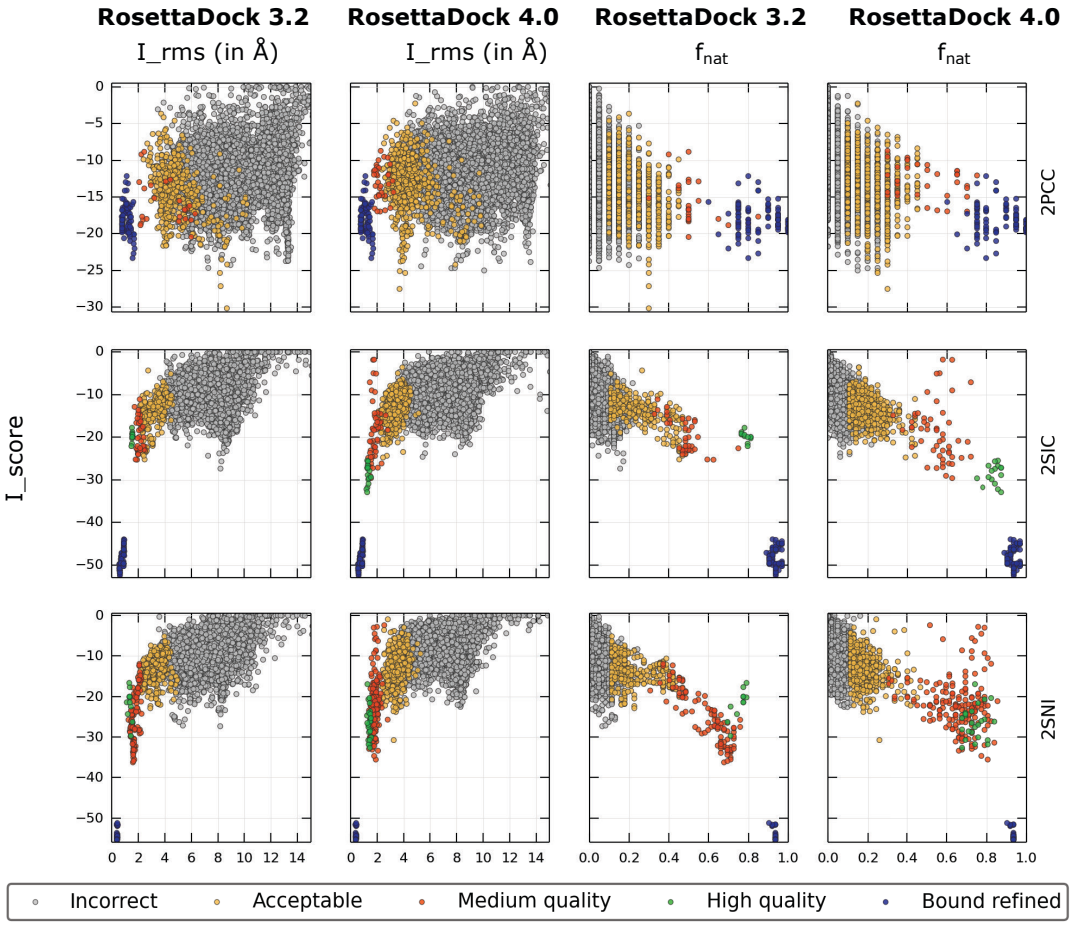
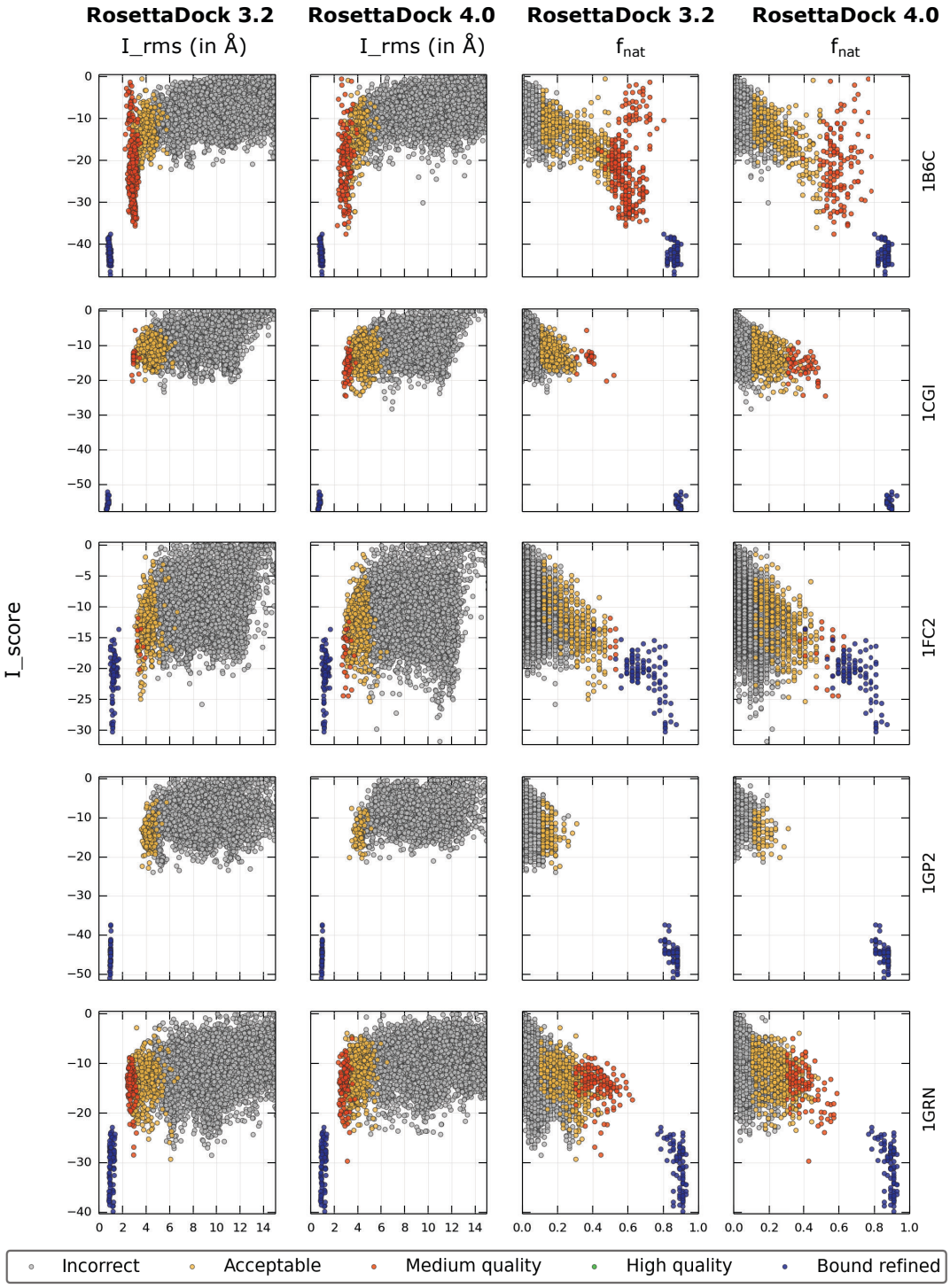
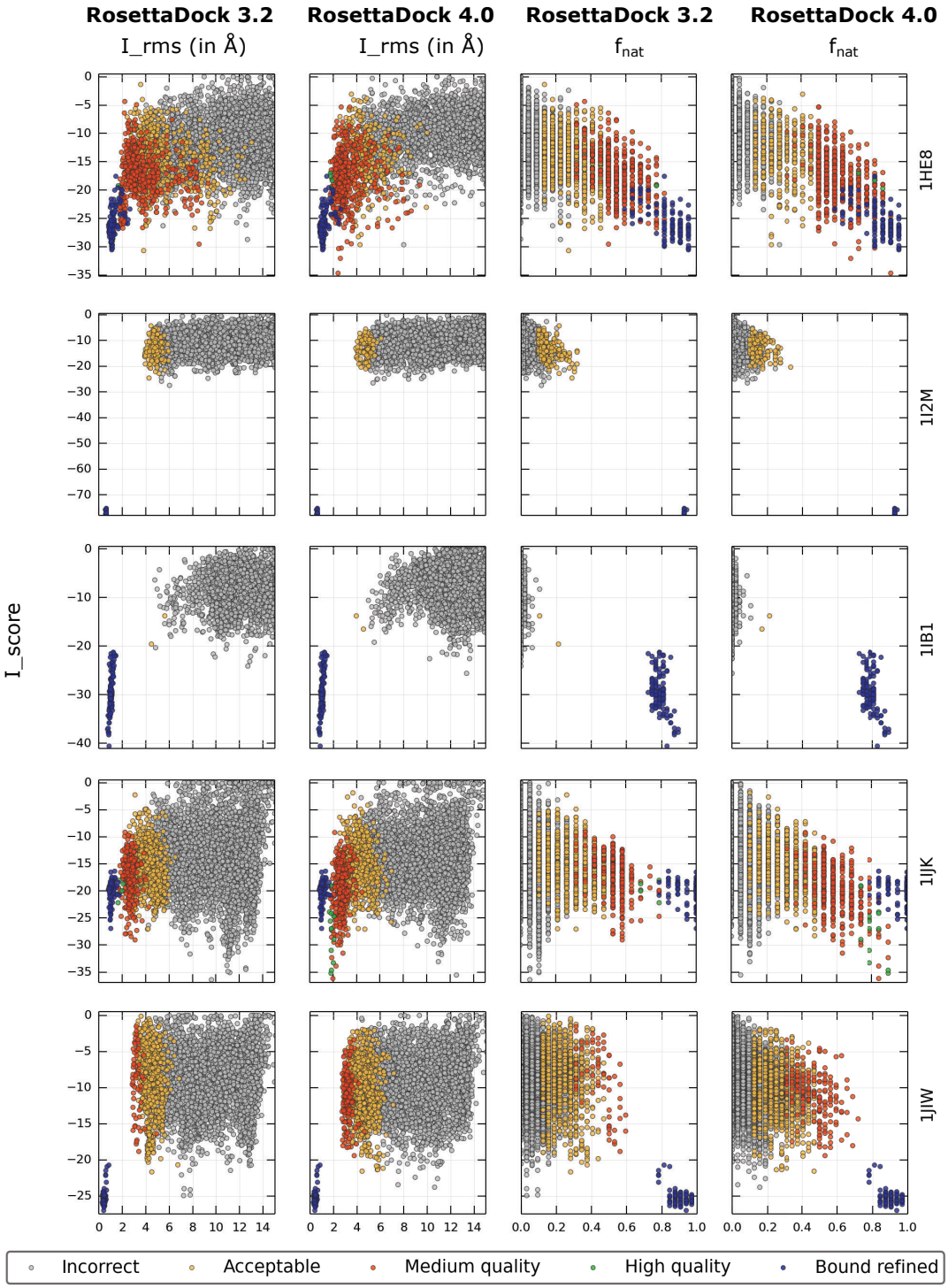
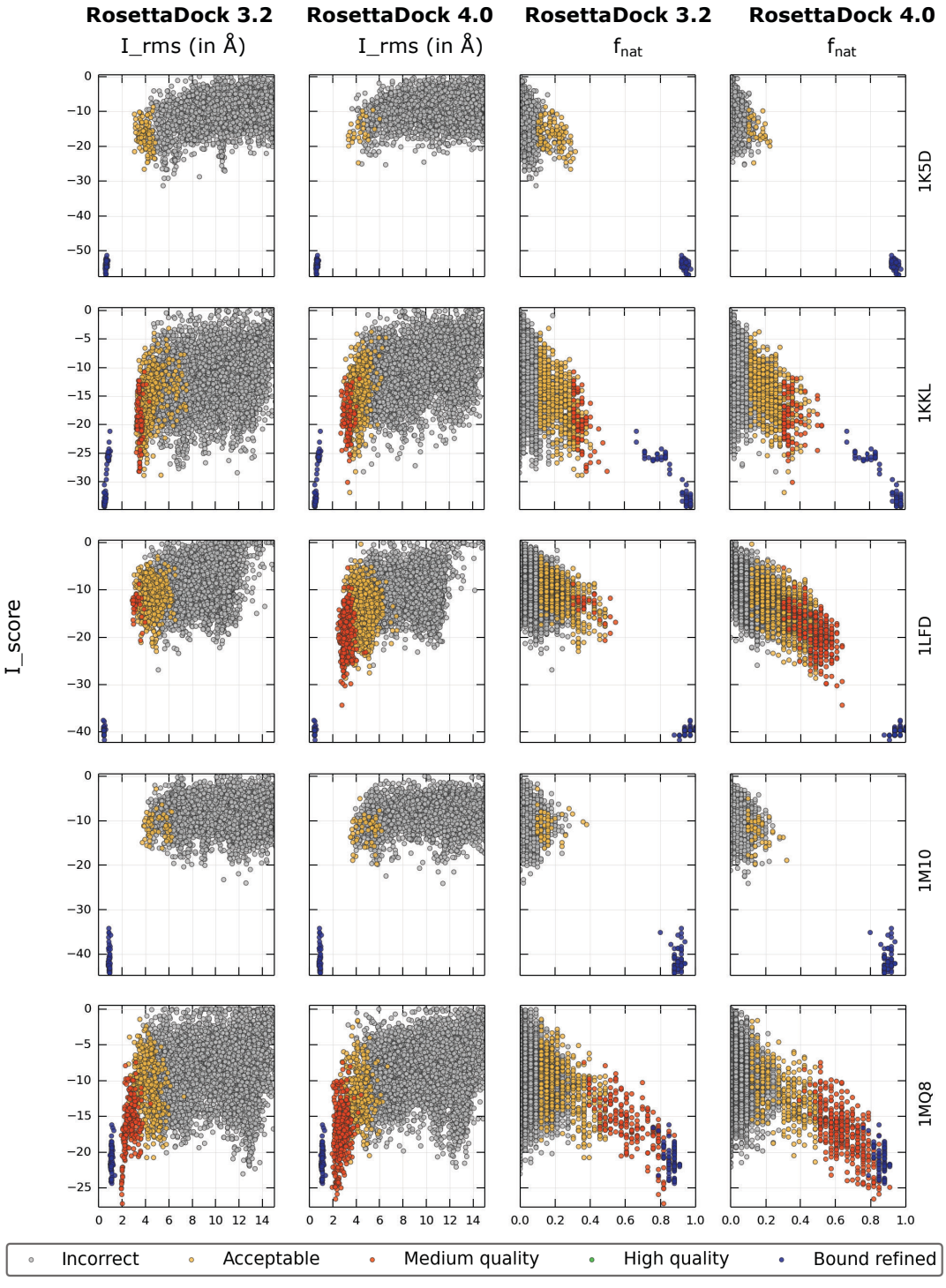
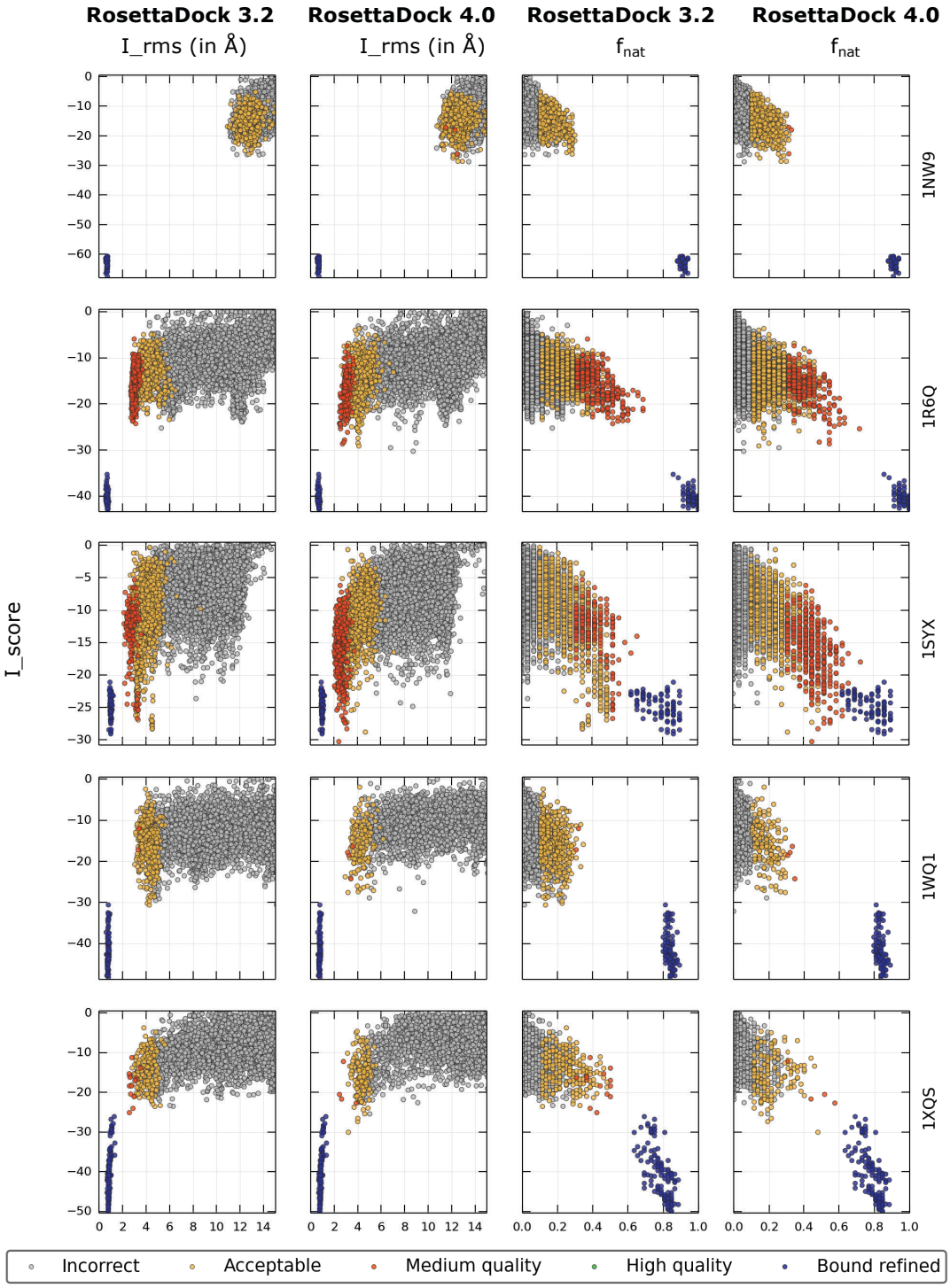


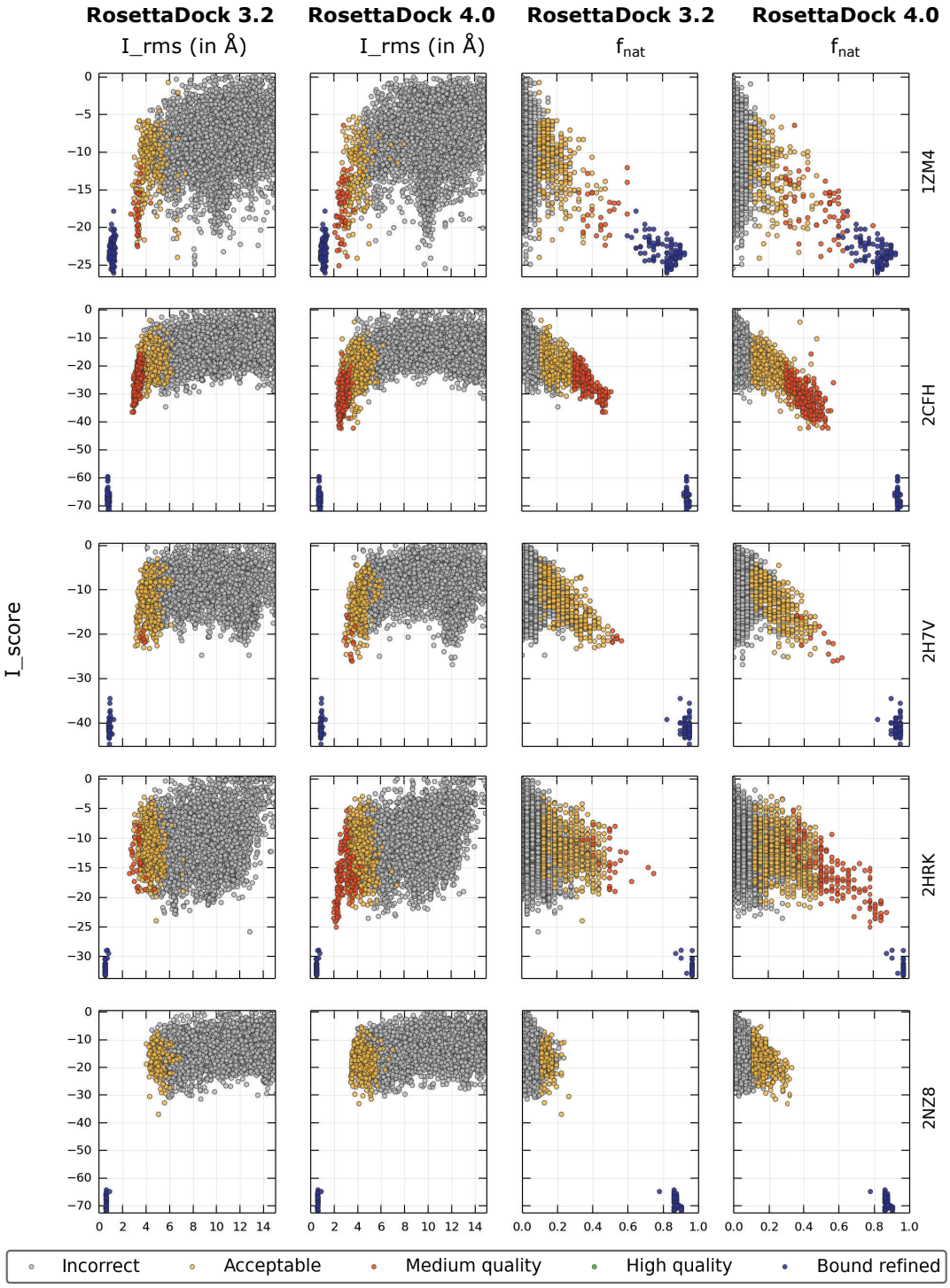
Figure A.4: Score versus RMSD plots & score versus f_{nat} plots after the full protocol for RosettaDock version 3.2 versus version 4.0 for rigid complexes.

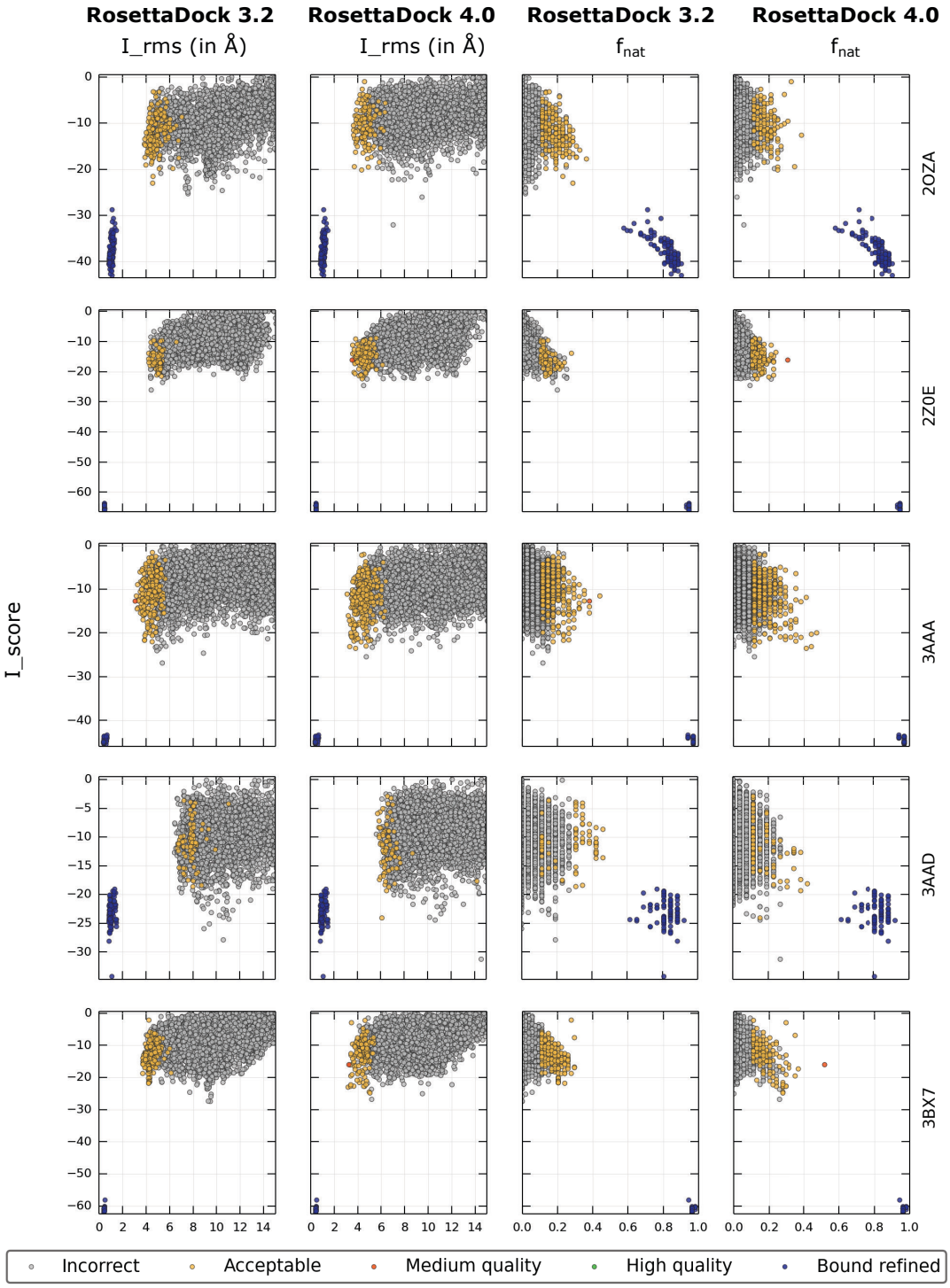

RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
I_score
1B6C
1CGI
1FC2
1GP2
1GRN
Incorrect
Acceptable
Medium quality
High quality
Bound refined

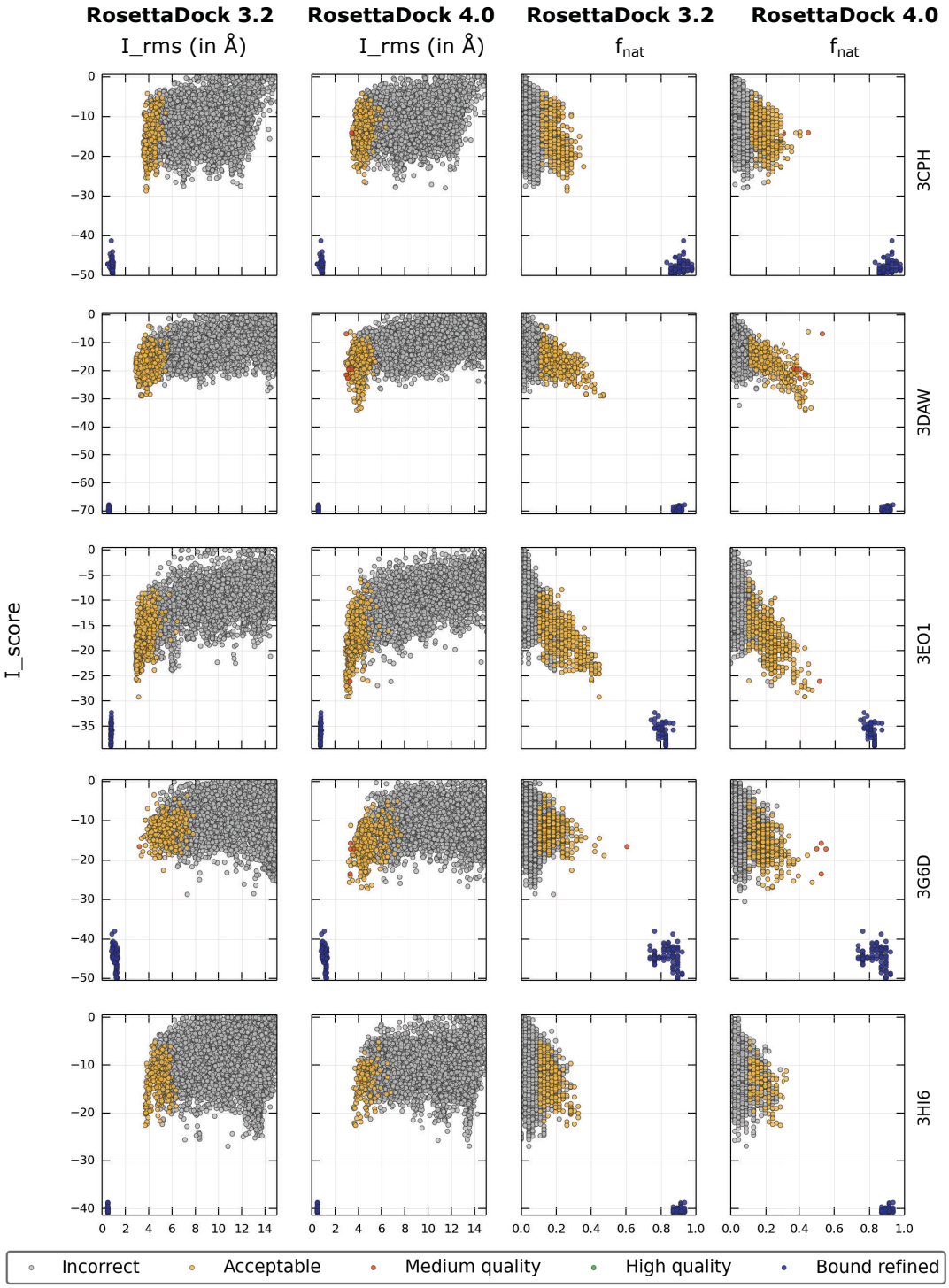

RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
I_score
1HE8
1I2M
1IB1
1IJK
1JIW
Incorrect
Acceptable
Medium quality
High quality
Bound refined

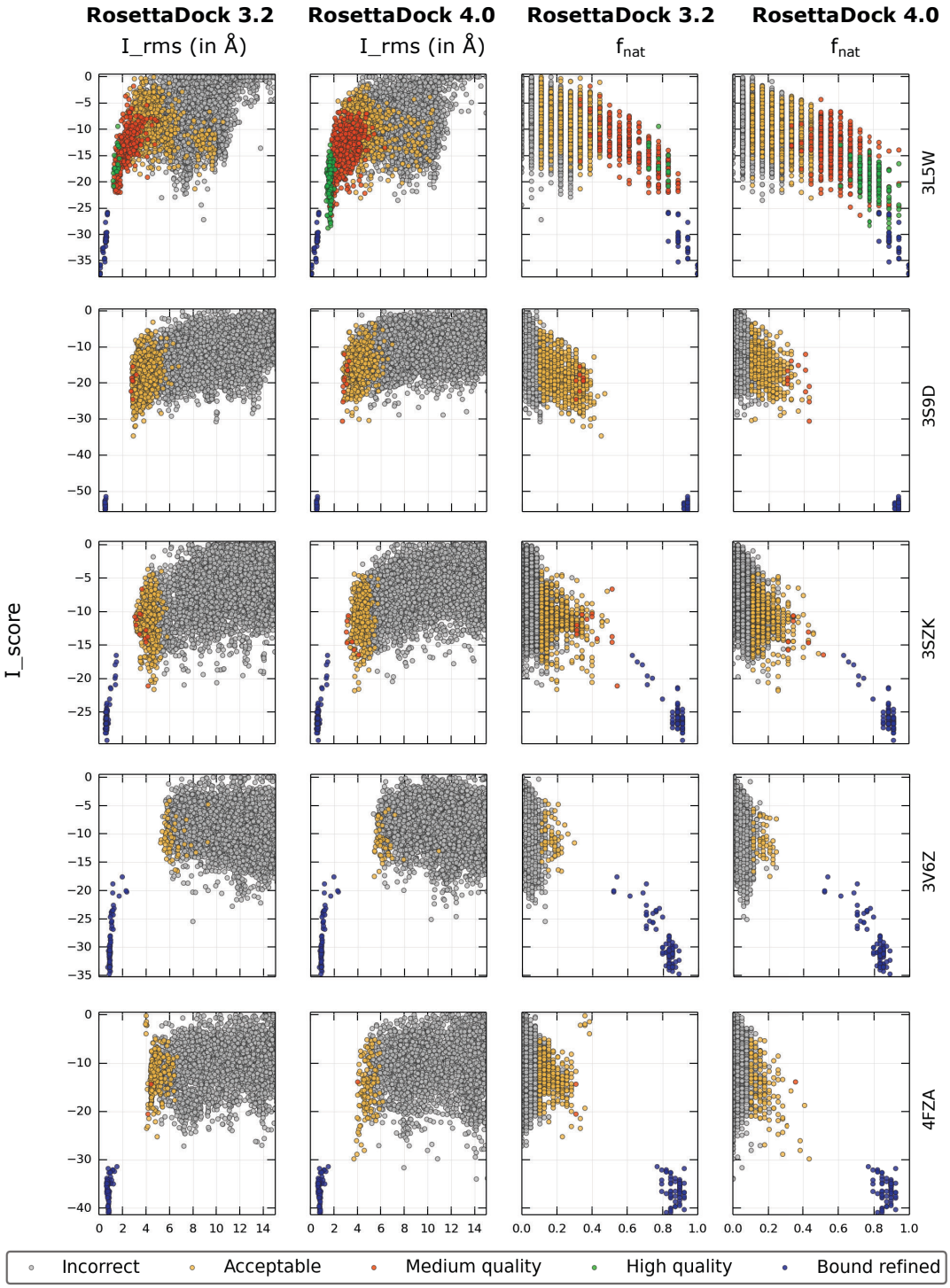

RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
I_score
1K5D
1KKL
1LFD
1M10
1MQ8
Incorrect
Acceptable
Medium quality
High quality
Bound refined


RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
I_score
1NW9
1R6Q
1SYX
1WQ1
1XQS
Incorrect
Acceptable
Medium quality
High quality
Bound refined


RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
I_score
1ZM4
2CFH
2H7V
2HRK
2NZ8
Incorrect
Acceptable
Medium quality
High quality
Bound refined


RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
I_score
2OZA
2Z0E
3AAA
3AAD
3BX7
Incorrect
Acceptable
Medium quality
High quality
Bound refined


RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
I_score
3CPH
3DAW
3EO1
3G6D
3HI6
Incorrect
Acceptable
Medium quality
High quality
Bound refined


RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
I_score
3L5W
3S9D
3SZK
3V6Z
4FZA
Incorrect
Acceptable
Medium quality
High quality
Bound refined

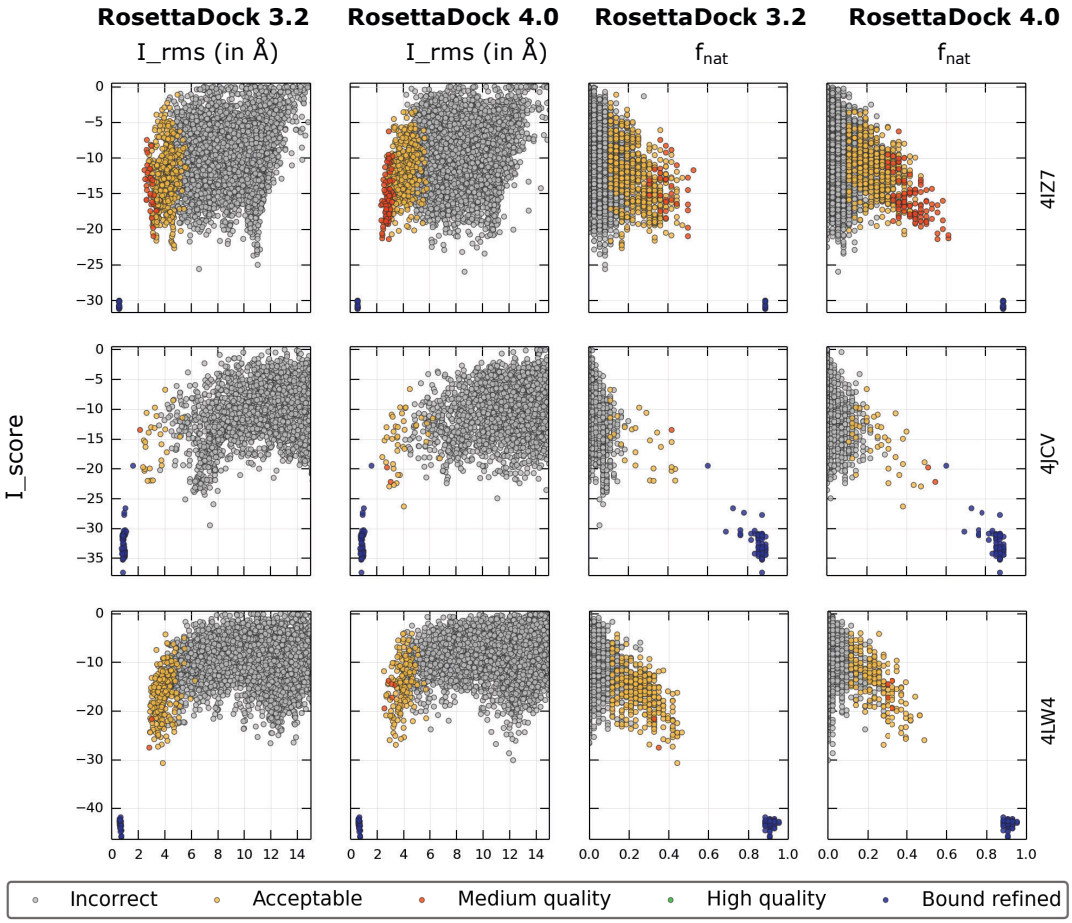
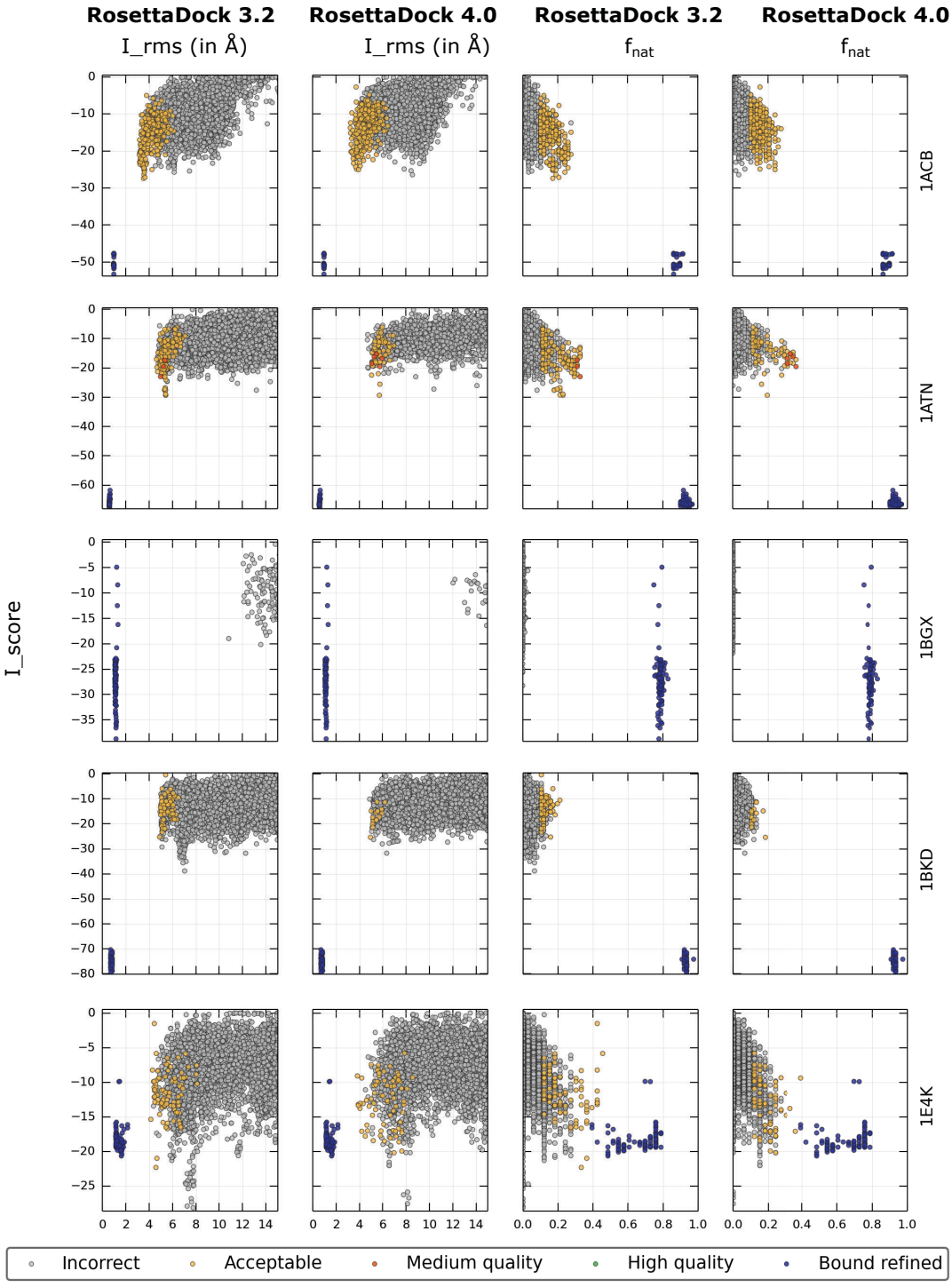
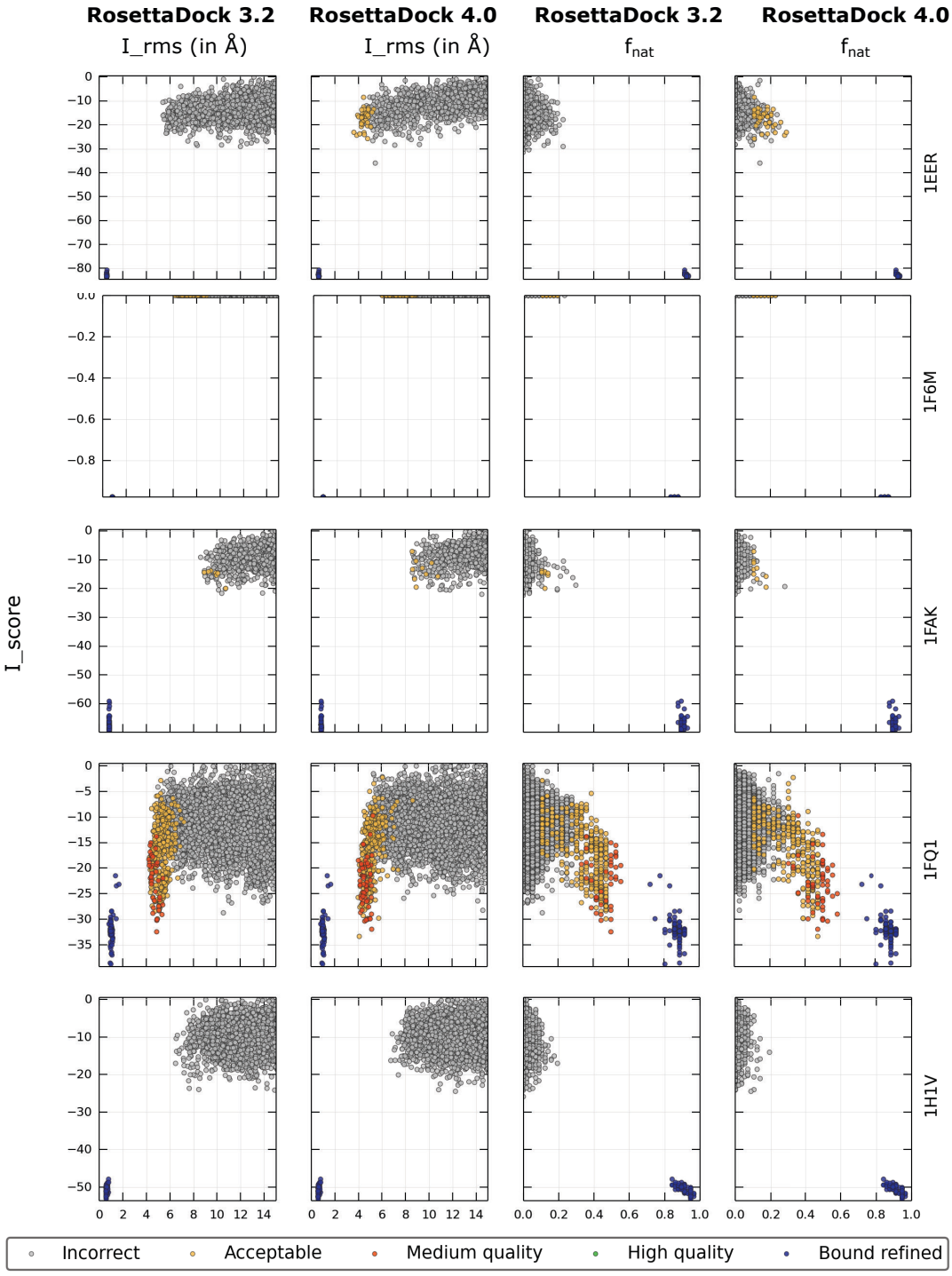
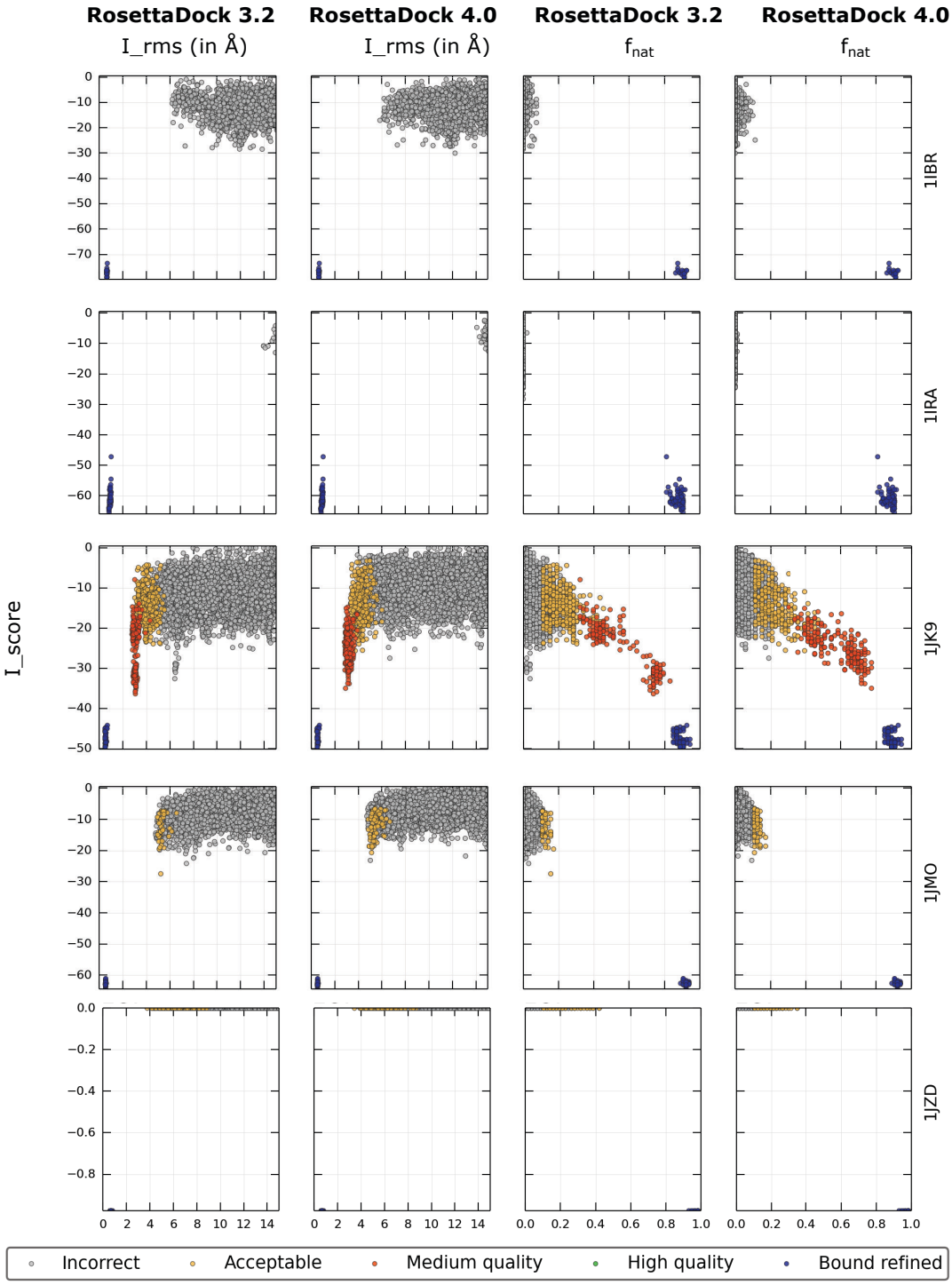
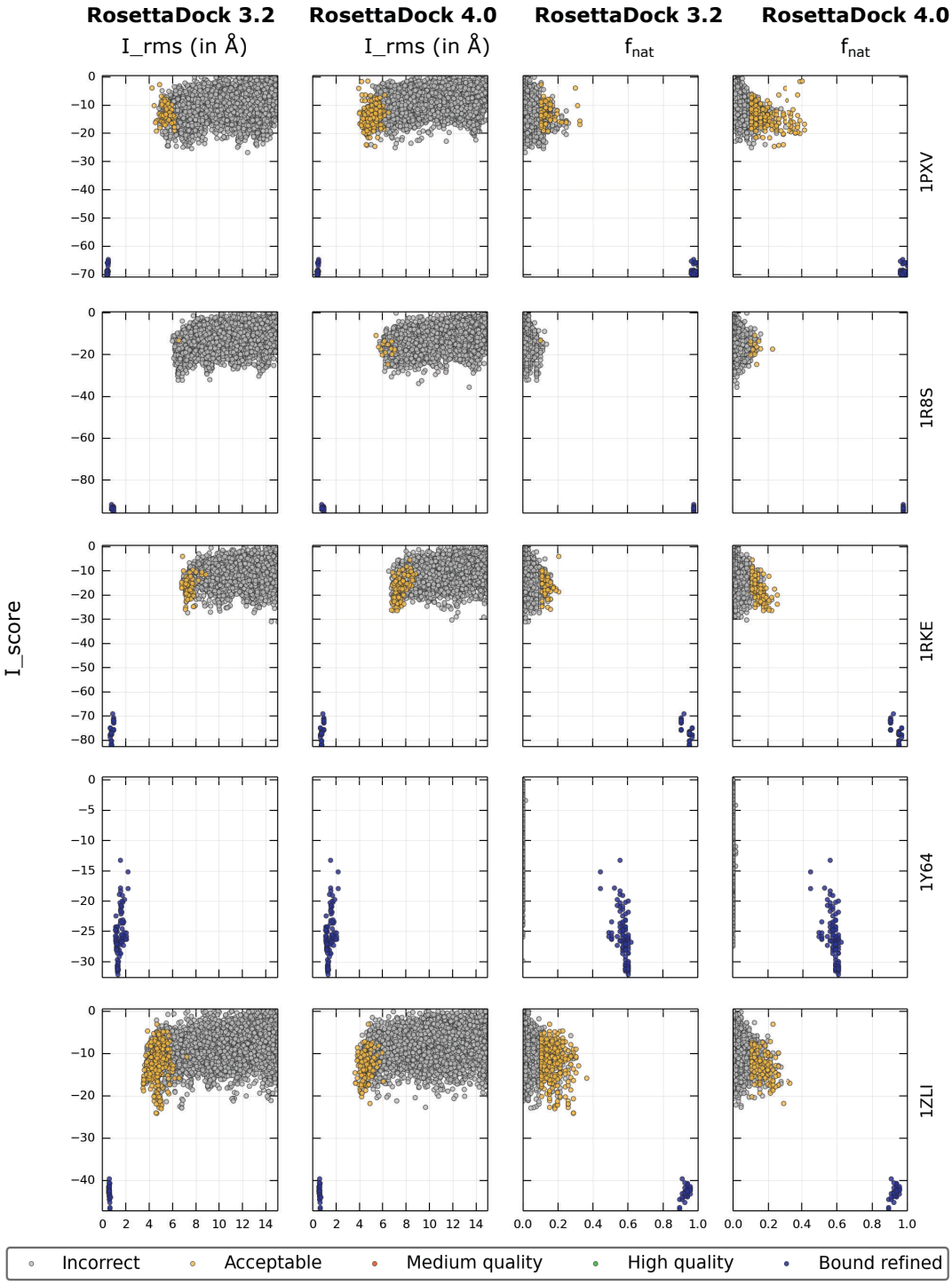


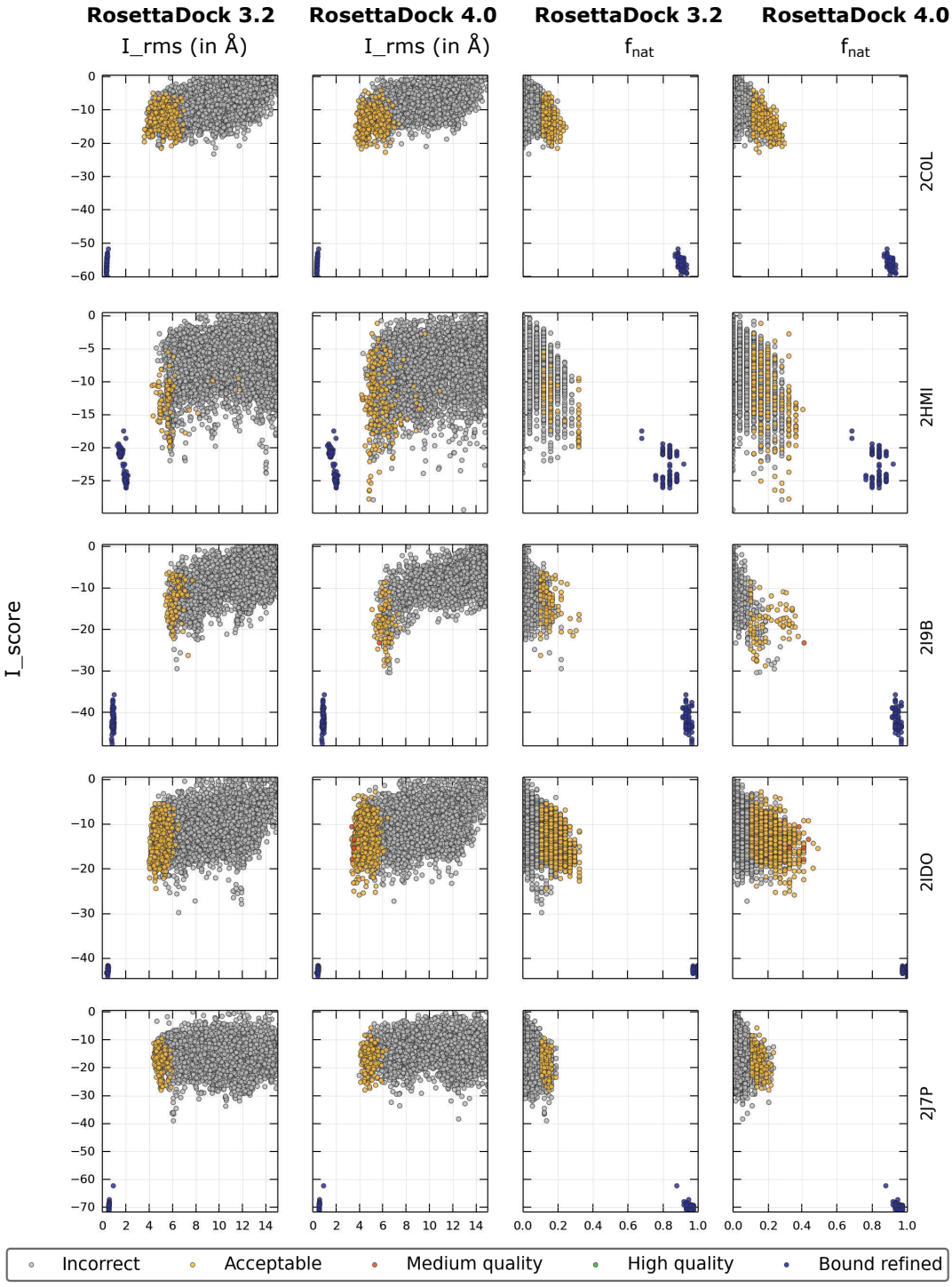
Figure A.5: Score versus RMSD plots & score versus f_{nat} plots after the full protocol for RosettaDock version 3.2 versus version 4.0 for medium-flexibility complexes.

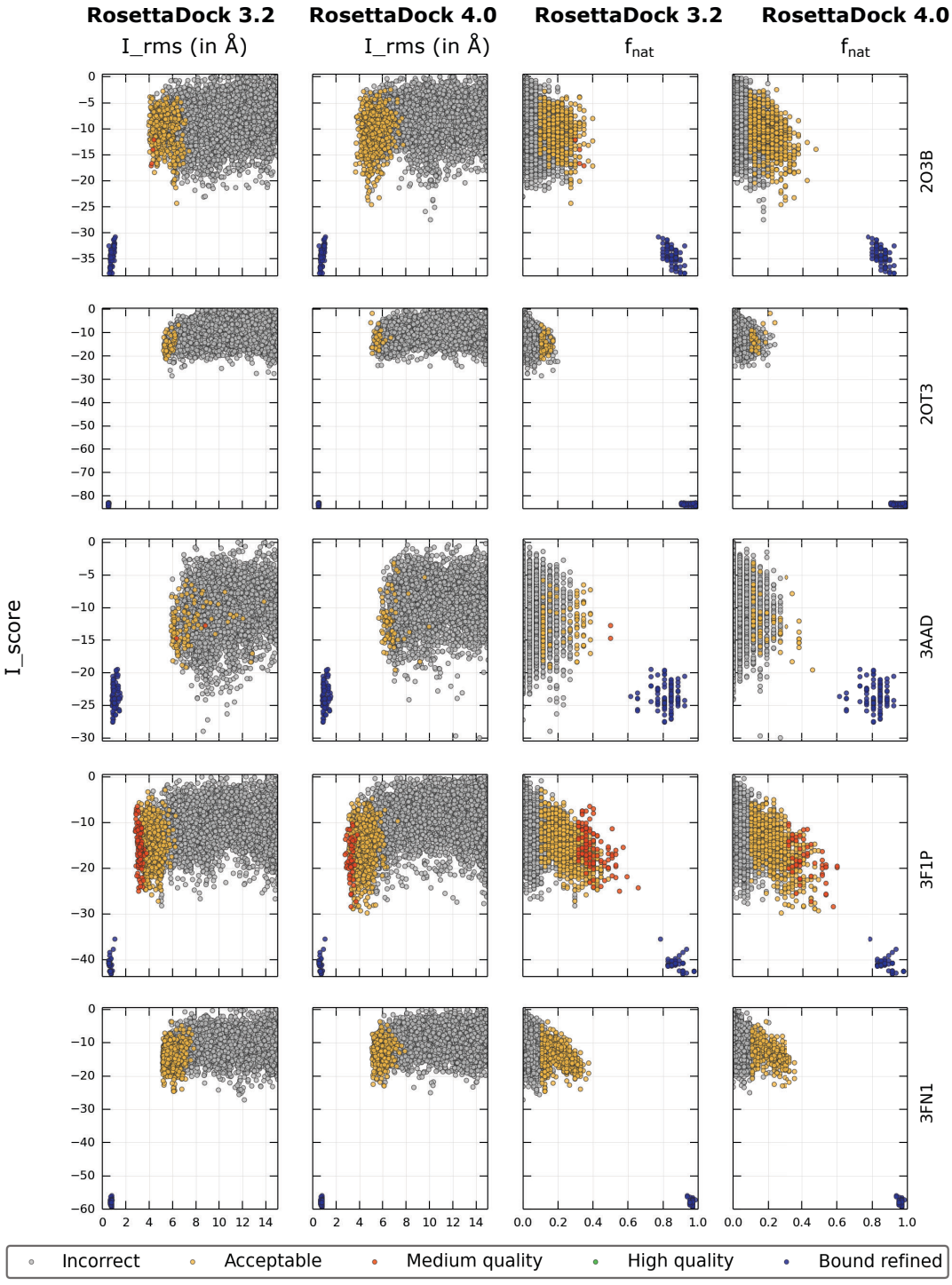

RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
I_score
1ACB
1ATN
1BGX
1BKD
1E4K
Incorrect
Acceptable
Medium quality
High quality
Bound refined


RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
I_score
1EER
1F6M
1FAK
1FQ1
1H1V
Incorrect
Acceptable
Medium quality
High quality
Bound refined


RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
1IBR
1IRA
1JK9
1JMO
1JZD
I_score
Incorrect
Acceptable
Medium quality
High quality
Bound refined


RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
I_score
1PXV
1R8S
1RKE
1Y64
1ZLI
Incorrect
Acceptable
Medium quality
High quality
Bound refined


RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
I_score
2C0L
2HMI
2I9B
2IDO
2J7P
Incorrect
Acceptable
Medium quality
High quality
Bound refined


RosettaDock 3.2
I_rms (in Å)
RosettaDock 4.0
I_rms (in Å)
RosettaDock 3.2
f_nat
RosettaDock 4.0
f_nat
I_score
2O3B
2OT3
3AAD
3F1P
3FN1
Incorrect
Acceptable
Medium quality
High quality
Bound refined

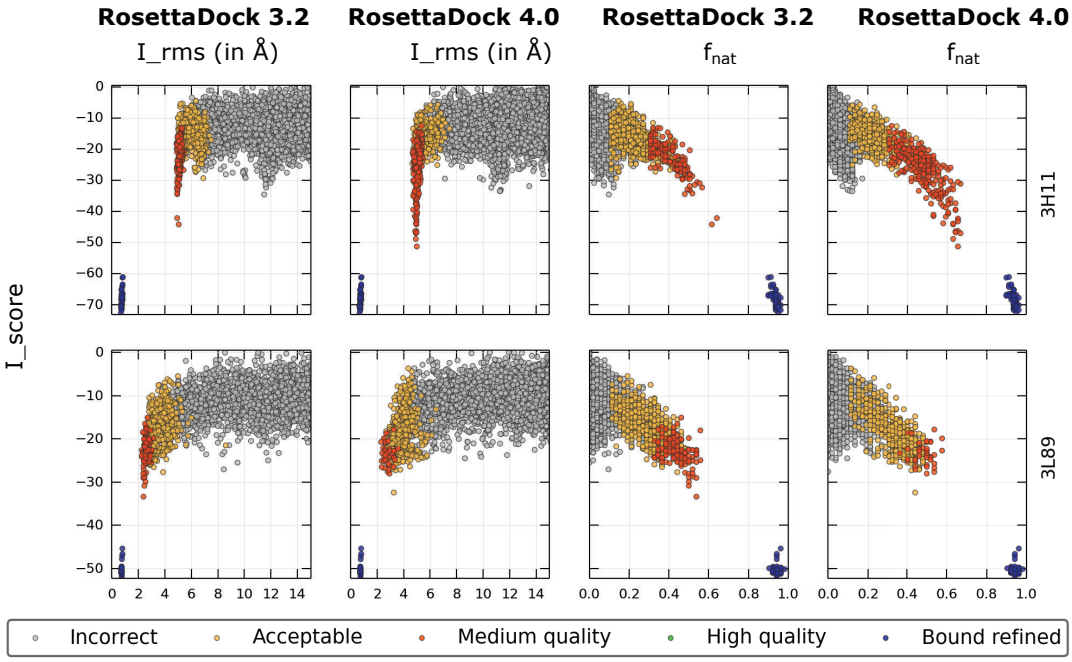


Figure A.6: Score versus RMSD plots & score versus f_{nat} plots after the full protocol for RosettaDock version 3.2 versus version 4.0 for flexible complexes.

cen_rms(in Å)
1ACB
1ATN
1BGX
1BKD
1E4K
1EER
1F6M
1FAK
1FQ1
1H1V
Motif Dock Score (without doping)
Motif Dock Score (with doping)
Incorrect
Acceptable
Medium quality
High quality
Bound re-dock

202

cen_rms(in Å)
1IBR
1IRA
1JK9
1JMO
1JZD
1PXV
1R8S
1RKE
1Y64
1ZLI
Motif Dock Score (without doping)
Motif Dock Score (with doping)
Incorrect
Acceptable
Medium quality
High quality
Bound re-dock

203

2C0L
2HMI
2I9B
2IDO
2J7P
2O3B
2OT3
3AAD
3F1P
3FN1
cen_rms(in Å)
Motif Dock Score (without doping)
Motif Dock Score (with doping)

204

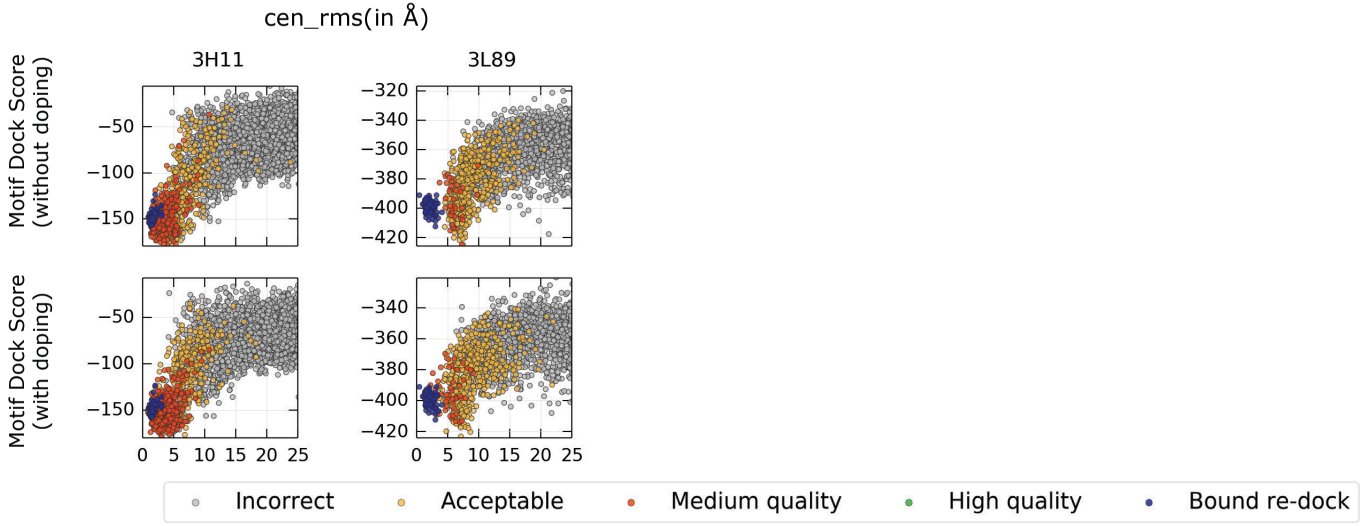
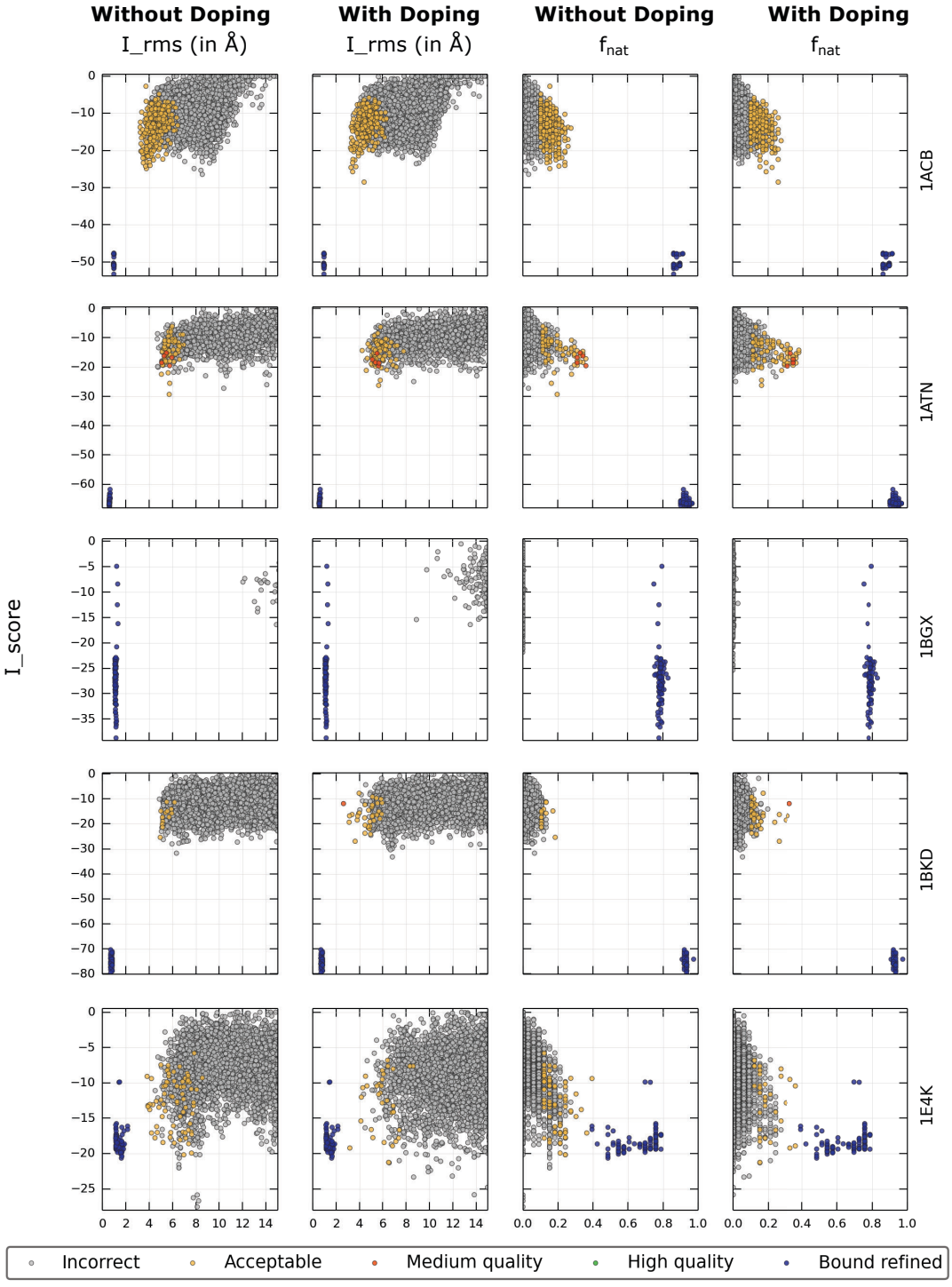
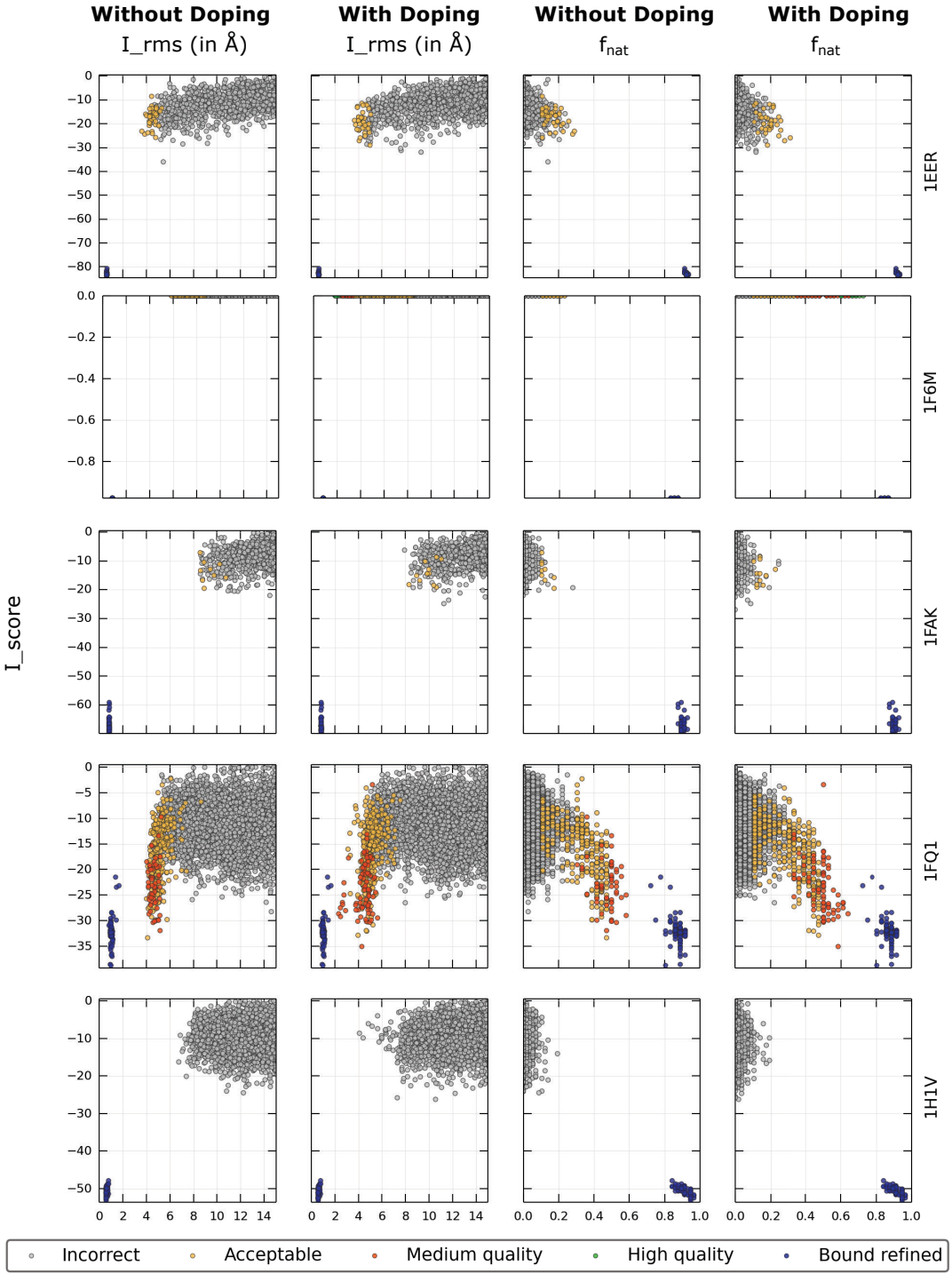
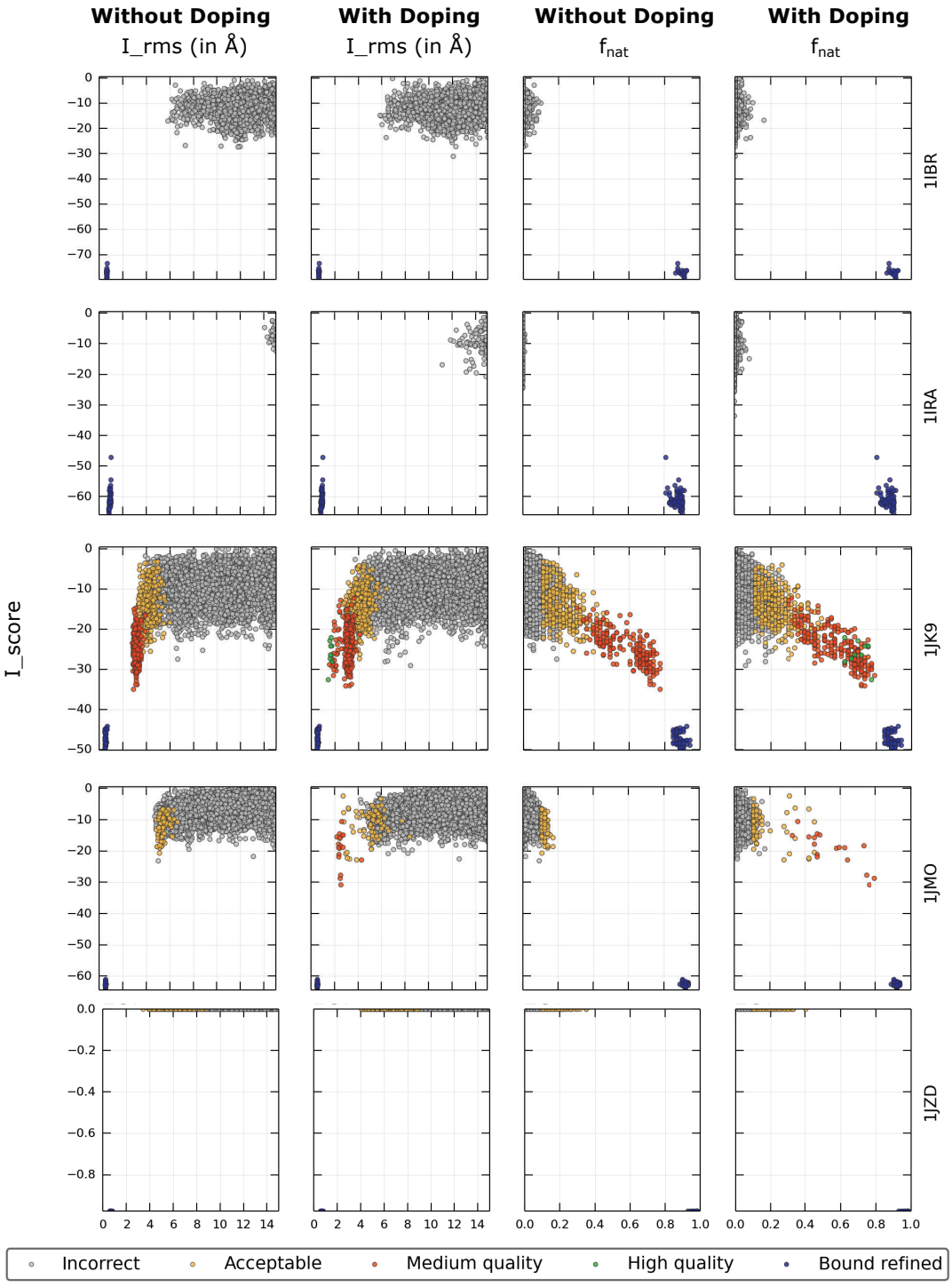
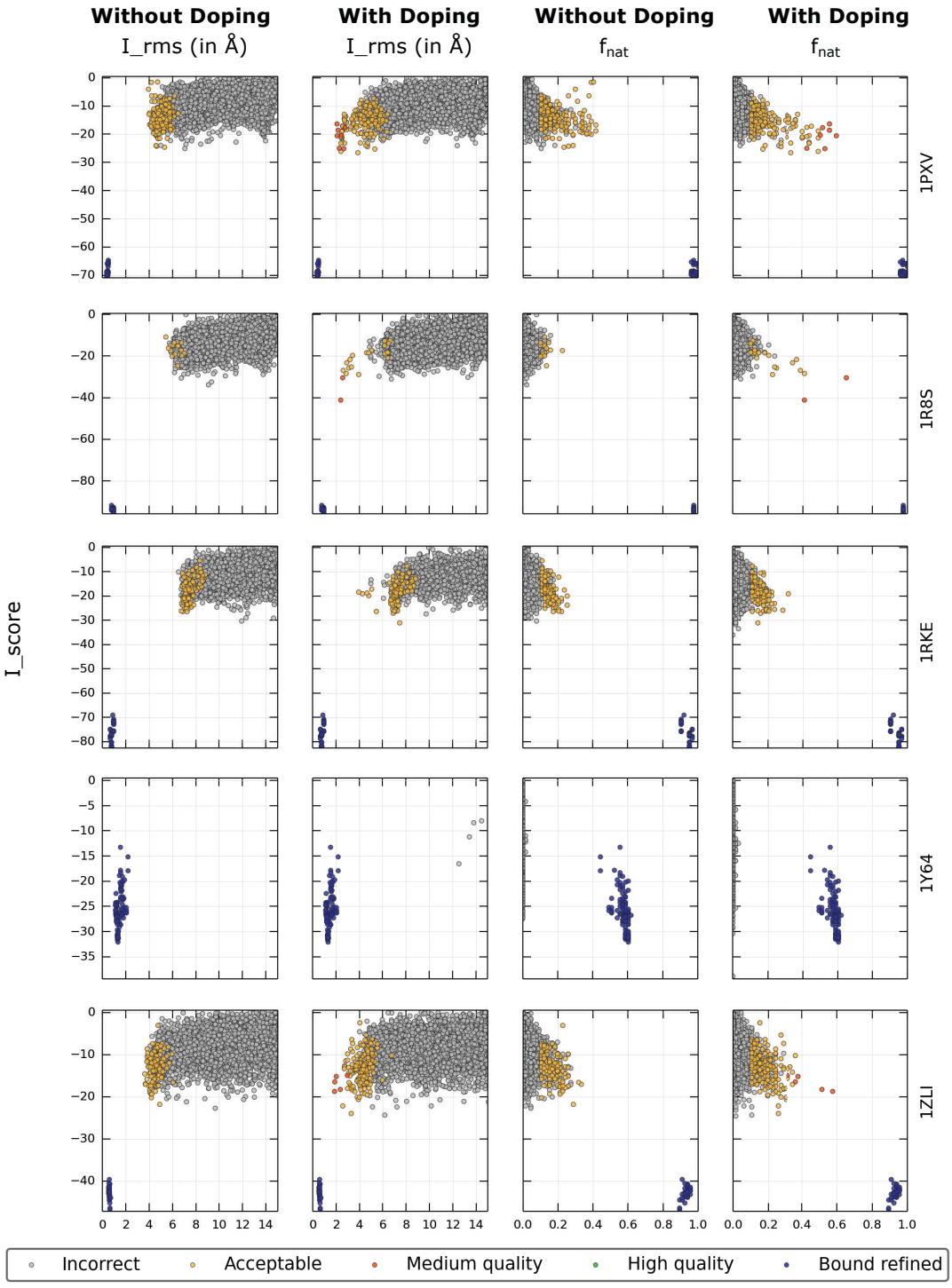


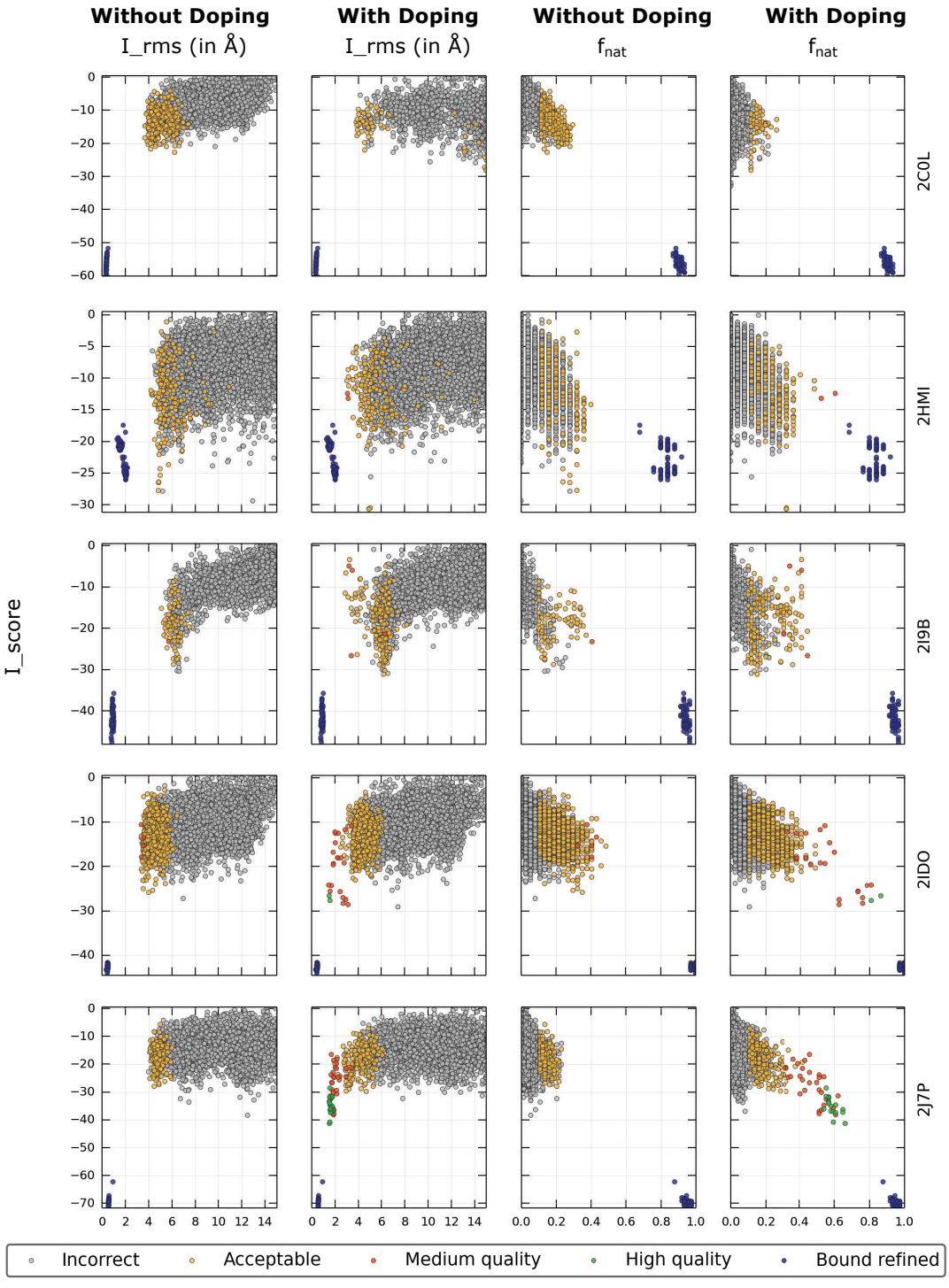
Figure A.7: Score versus RMSD plots in the low-resolution stage for flexible complexes with and without ensemble doping.

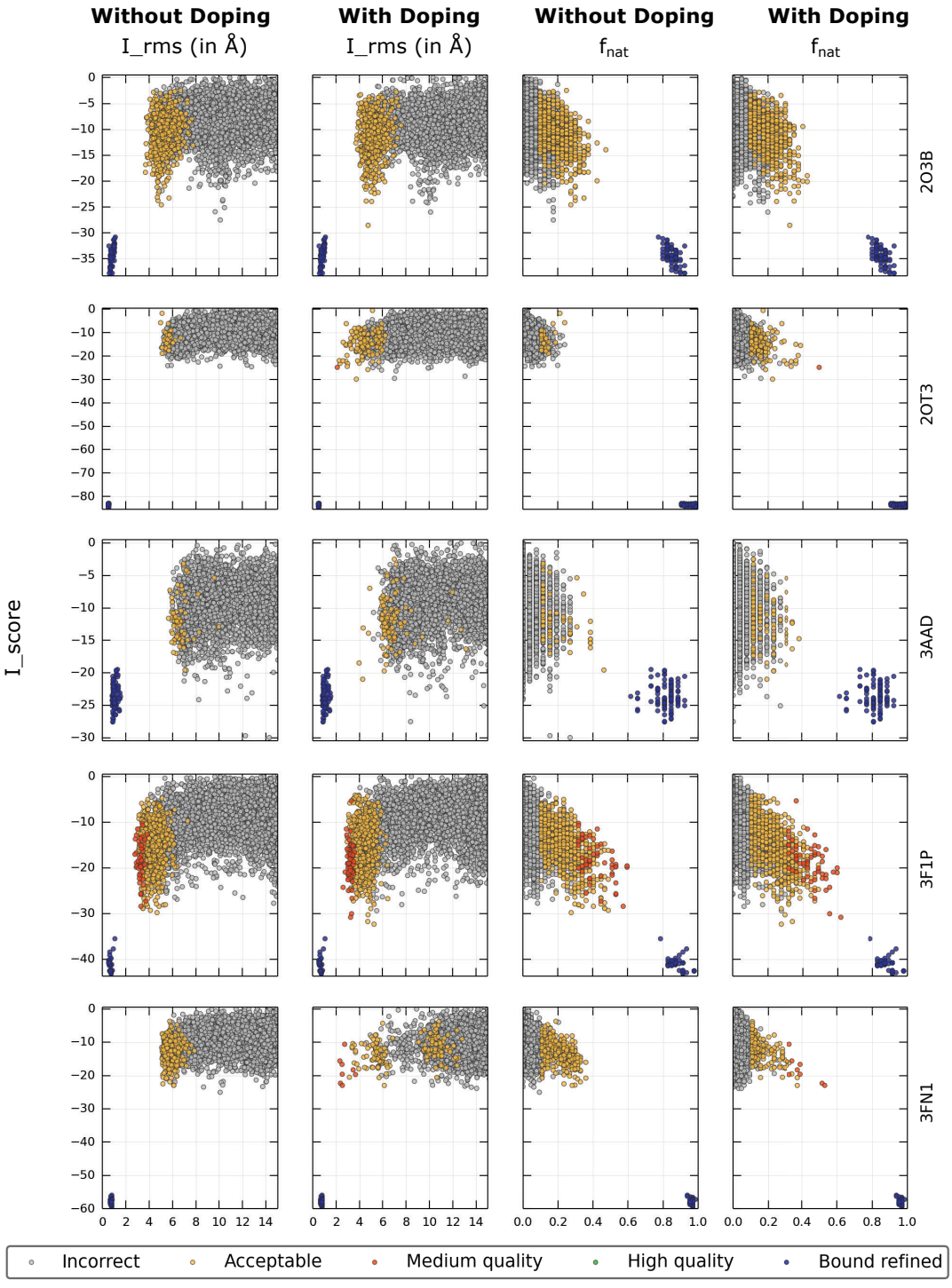

Without Doping
I_rms (in Å)
With Doping
I_rms (in Å)
Without Doping
f_nat
With Doping
f_nat
1ACB
1ATN
1BGX
1BKD
1E4K
I_score
Incorrect
Acceptable
Medium quality
High quality
Bound refined


Without Doping
I_rms (in Å)
With Doping
I_rms (in Å)
Without Doping
f_nat
With Doping
f_nat
1EER
1F6M
1FAK
1FQ1
1H1V
I_score
Incorrect
Acceptable
Medium quality
High quality
Bound refined


Without Doping
I_rms (in Å)
With Doping
I_rms (in Å)
Without Doping
f_nat
With Doping
f_nat
I_score
1IBR
1IRA
1JK9
1JMO
1JZD
Incorrect
Acceptable
Medium quality
High quality
Bound refined


Without Doping
I_rms (in Å)
With Doping
I_rms (in Å)
Without Doping
f_nat
With Doping
f_nat
I_score
1PXV
1R8S
1RKE
1Y64
1ZLI
Incorrect
Acceptable
Medium quality
High quality
Bound refined


Without Doping
I_rms (in Å)
With Doping
I_rms (in Å)
Without Doping
f_nat
With Doping
f_nat
I_score
2C0L
2HMI
2I9B
2IDO
2J7P
Incorrect
Acceptable
Medium quality
High quality
Bound refined


Without Doping
I_rms (in Å)
With Doping
I_rms (in Å)
Without Doping
f_nat
With Doping
f_nat
I_score
2O3B
2OT3
3AAD
3F1P
3FN1
Incorrect
Acceptable
Medium quality
High quality
Bound refined

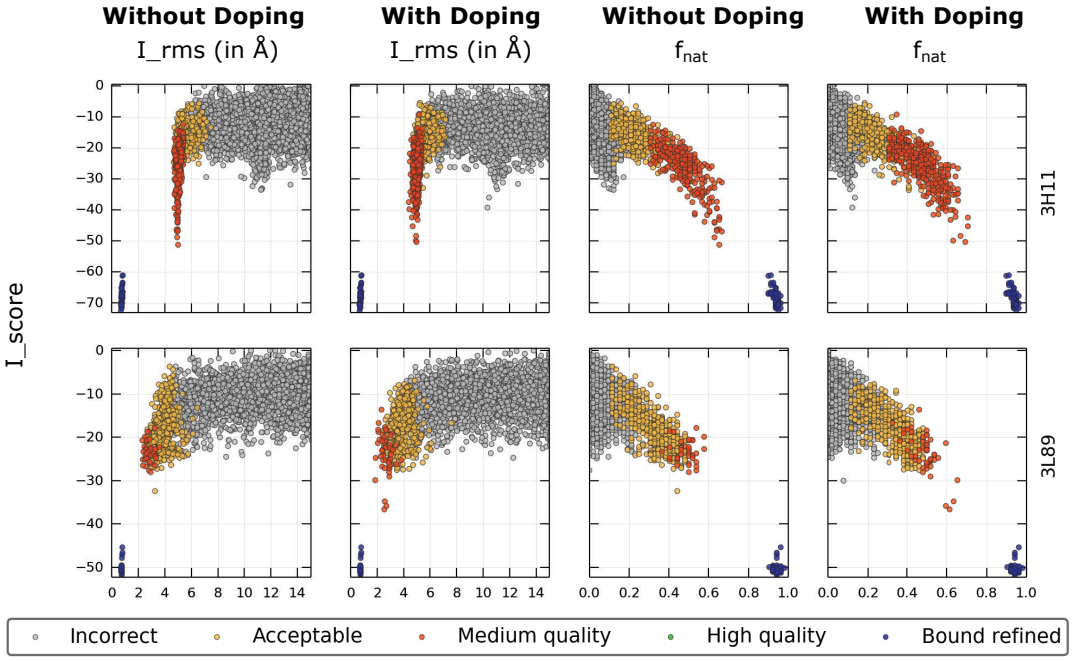


Figure A.8: Score versus RMSD plots & score versus f_{nat} plots after the full protocol for RosettaDock version 3.2 versus version 4.0 for flexible complexes without and with ensemble doping.

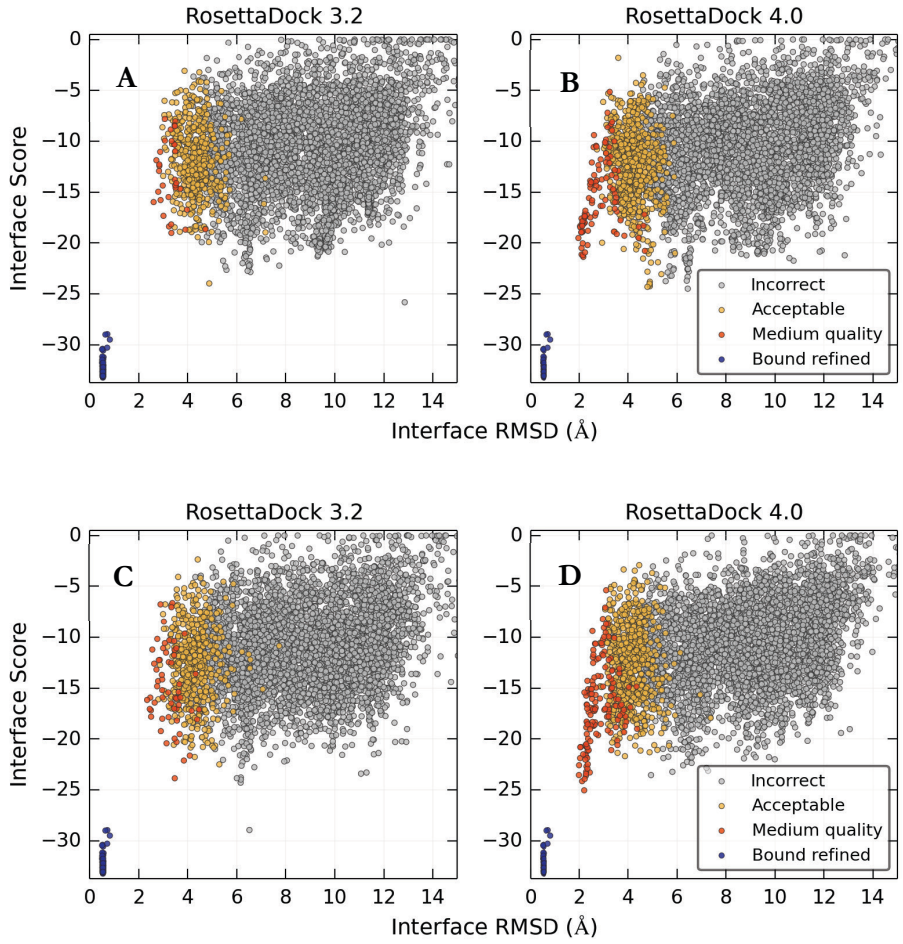


Figure A.9: Interface score versus interface RMSD plots for docking simulation of glutamyl-tRNA synthetase–GU4 nucleic-binding protein 1 complex with (A) RosettaDock 3.2 and ensembles with 1 receptor and 10 ligand conformations, (B) RosettaDock 4.0 and ensembles with 1 receptor and 10 ligand conformations, (C) RosettaDock 3.2 and ensembles with 100 conformations each of the receptor and the ligand, and (D) RosettaDock 4.0 and ensembles with 100 conformations each of the receptor and the ligand. (B) and (D) are enriched in medium-quality docked models as compared to (A) and (C), respectively. (C) has a deeper funnel than (A) owing to the inclusion of structures generated by Backrub, which produces loop motions that mimic the unbound-bound conformational change. (D) has both a deep funnel and enhanced sampling.

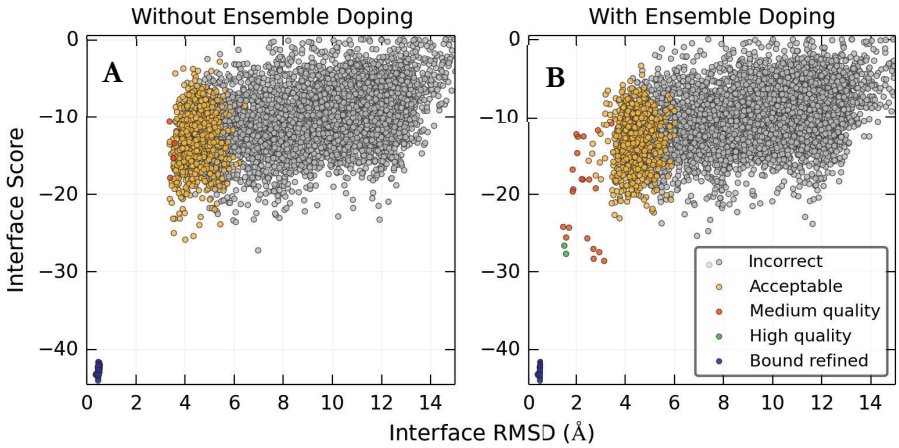


Figure A.10: Improvement in docking performance of RosettaDock 4.0 by doping the ensemble with near-bound decoys for Pol III-ε–Hot complex. (A) Score versus RMSD plot of runs with backbone conformations generated using NMA, Backrub and Relax protocols do not have medium- or high-quality docked structures. (B) 10% doping with near-bound conformations leads to deep docking funnels with high-quality structures.

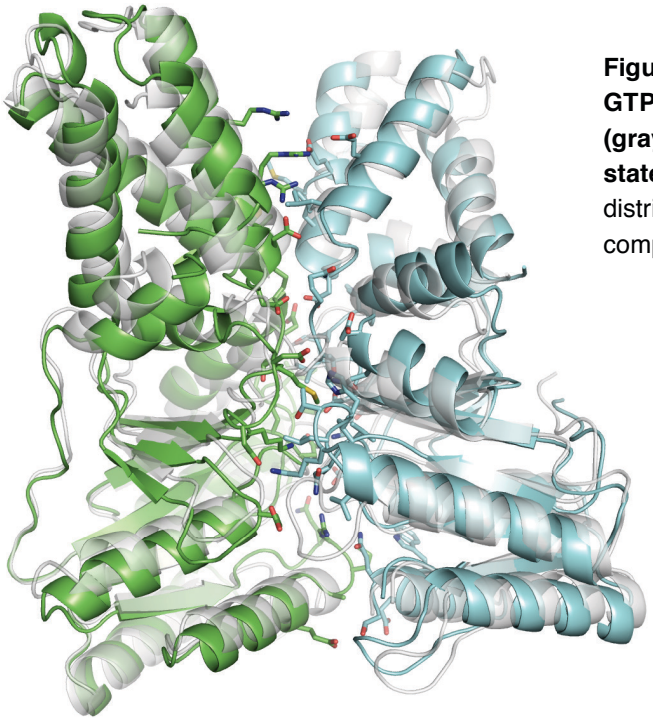


Figure A.11: The structure of the SRP GTPase–FtsY complex in the unbound (gray) and the bound (cyan/green) states. The bound-unbound motion is distributed across a large interface in the complex.

Appendix B

Modeling of symmetric homomeric complexes

Table B.1: Benchmark of 43 symmetric homomeric complexes

Target	Symmetry	Resolution (Å)	Type	Subunit size (residues)	Earliest REVDAT
3urr	C2	1.4	transferase	153	21-Dec-11
1wpn*	C2	1.3	hydrolase	188	23-Nov-04
2nlv	C2	1.3	XisI-like (unknown function)	112	5-Dec-06
3m1z*	C2	1.42	lyase	228	16-Jun-10
4zo2	C2	1.09	hydrolase	294	15-Jul-15
1sg4*	C3	1.3	isomerase	260	18-Jan-05
4co0	C3	1.4	signaling protein	112	28-May-14
4d7y*	C3	1.44	signaling protein	146	28-Jan-15
4xla	C3	1.47	viral protein	600	20-Jan-16
5i6n	C3	1.22	oxidoreductase	332	13-Jul-16
1p5b	C4	1.35	oxidoreductase	380	28-Oct-03
2o6n*	C4	1.1	de novo protein	35	23-Oct-07

2v9n*	C4	1.4	lyase	274	15-Jan-08
4z0g	C4	1.25	oxidoreductase	413	25-Nov-15
3v9o*	C4	1.45	lyase	143	25-Jan-12
1xb9*	C5	1.9	chaperone	114	21-Dec-04
4avs	C5	1.4	sugar binding protein	204	19-Jun-13
4u62	C5	1.55	viral protein	280	5-Aug-15
5a12	C5	1.4	oxidoreductase	242	2-Sep-15
5lzh	C5	1.13	toxin	103	31-May-17
3h47*	C6	1.9	viral protein	231	23-Jun-09
4ox6*	C6	1.34	structural protein	127	27-Aug-14
2xf7	C6	1.61	viral protein	51	11-Aug-10
1nlf	C6	1.95	replication	279	29-Apr-03
4w64	C6	1.55	Hcp1 protein (Unknown function)	171	1-Jul-15
4owk	C7	2.0	toxin	138	28-May-14
1h64	C7	1.9	Sm-like protein	75	19-Dec-02
4f87	C8	1.4	antimicrobial protein; viral protein	72	25-Jul-12
3b8o	C8	2.4	biosynthetic protein	265	22-Jan-08
3zqo	C9	1.68	DNA binding protein	72	28-Dec-11
3p9a	C9	1.75	DNA binding protein	162	9-May-12
1orr	D2	1.5	isomerase	347	26-Aug-03
1zjz	D2	1.1	oxidoreductase	251	21-Jun-05
2bv4	D2	1.0	lectin	113	25-May-06
4oqc	D2	1.3	oxygen binding	302	24-Dec-14
2vqr	D2	1.42	hydrolase	543	30-Sep-08
3v4f	D3	1.39	signaling protein	166	8-Aug-12

2bhq	D3	1.4	oxidoreductase	516	9-Mar-06
3qns	D3	1.4	oxidoreductase	353	27-Apr-11
2j5g	D3	1.46	hydrolase	263	16-Jan-07
1gxu	D3	1.27	phosphatase	91	12-Sep-02
2r8e	D4	1.4	hydrolase	188	23-Sep-08
3r1m	D4	1.5	metal binding protein	385	12-Oct-11

* chosen for testing flexible-backbone strategies

Table B.2: $RMSD_{C\alpha}$ (Å) from native monomer of five homology-modeled monomers obtained from sequence using Robetta

PDB ID	Symmetry	Model 1	Model 2	Model 3	Model 4	Model 5
3urr	C2	0.430	0.430	1.243	0.555	0.503
1wpn	C2	0.080	0.080	0.067	0.065	0.072
2nlv	C2	1.312	1.335	1.299	15.618	6.936
3m1z	C2	0.054	0.054	0.054	0.051	0.053
4zo2	C2	1.496	1.530	1.311	1.435	1.440
1sg4	C3	0.199	0.199	0.194	0.180	0.183
4co0	C3	0.125	0.125	0.167	0.104	0.123
4d7y	C3	0.944	0.944	0.743	0.894	0.967
4xla	C3	0.315	0.316	0.322	0.318	0.315
5i6n	C3	0.362	0.358	0.343	0.352	0.355
1p5b	C4	0.155	0.155	0.141	0.150	0.146
2o6n	C4	0.159	0.159	13.327	12.365	9.884
2v9n	C4	0.321	0.333	0.332	0.327	0.327
4z0g	C4	0.123	0.123	0.109	0.101	0.112
3v9o	C4	1.064	1.064	0.877	0.941	1.013
1xb9	C5	0.199	0.199	0.347	0.323	5.444
4avs	C5	0.373	0.393	0.370	0.384	0.362
4u62	C5	0.123	0.123	0.125	0.122	0.138
5a12	C5	0.639	0.689	0.654	0.669	0.673
5lzh	C5	0.357	0.361	0.379	0.339	0.351
3h47	C6	0.178	0.178	0.177	0.090	14.601
4ox6	C6	0.123	0.123	0.107	0.169	0.155
2xf7	C6	1.539	1.539	3.235	8.779	10.379
1nlf	C6	0.236	0.236	0.280	0.194	0.207
4w64	C6	0.982	0.982	0.908	0.894	0.726
4owk	C7	0.976	0.976	0.905	0.847	1.085
1h64	C7	0.132	0.132	0.091	0.104	0.122
4f87	C8	3.195	3.195	5.288	6.267	6.388
3b8o	C8	16.076	16.076	9.756	12.476	14.830
3zqo	C9	13.527	13.527	10.004	10.721	7.345
3p9a	C9	17.069	17.372	12.667	17.078	17.078

1orr	D2	0.912	0.912	0.839	1.096	0.710
1zjz	D2	0.421	0.440	0.456	0.464	0.453
2bv4	D2	0.417	0.424	0.435	0.413	0.393
4oqc	D2	0.156	0.156	0.147	0.150	0.163
2vqr	D2	2.235	2.076	2.202	1.986	1.986
3v4f	D3	0.385	0.382	0.396	0.382	0.386
2bhq	D3	0.479	0.474	0.489	0.471	0.479
3qns	D3	0.196	0.196	0.208	0.208	0.208
2j5g	D3	0.156	0.178	0.175	0.186	0.186
1gxu	D3	6.593	6.835	6.782	6.046	6.046
2r8e	D4	0.372	0.348	0.367	0.385	0.385
3r1m	D4	0.067	0.067	0.069	20.772	23.647

Table B.3: Performance of Rosetta SymDock vs. Rosetta SymDock2 across a 43-target benchmark set.

		Rosetta SymDock2				Rosetta SymDock			
PDBID	Symmetry	⟨E1%⟩	Coarse-grained ⟨N5⟩ [P(⟨N5⟩ ≥ 3)]	Full protocol ⟨N5⟩ [P(⟨N5⟩ ≥ 3)]	Full protocol ⟨N10⟩ [P(⟨N10⟩ ≥ 1)]	⟨E1%⟩	Coarse-grained ⟨N5⟩ [P(⟨N5⟩ ≥ 3)]	Full protocol ⟨N5⟩ [P(⟨N5⟩ ≥ 3)]	Full protocol ⟨N10⟩ [P(⟨N10⟩ ≥ 1)]
1WPN	C2	37.1 ± 4.3	5.0 ± 0.0 [1.00]	3.8 ± 1.0 [0.90]	6.9 ± 1.5 [1.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.2 [0.00]	0.5 ± 0.8 [0.37]
2NLV	C2	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.3 ± 0.7 [0.02]	1.7 ± 1.3 [0.81]	12.5 ± 4.7	0.0 ± 0.1 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
3M1Z	C2	38.7 ± 3.5	5.0 ± 0.0 [1.00]	5.0 ± 0.0 [1.00]	10.0 ± 0.0 [1.00]	12.4 ± 2.5	3.9 ± 1.1 [0.89]	4.1 ± 0.9 [0.95]	7.7 ± 1.4 [1.00]
3URR	C2	4.2 ± 2.5	0.7 ± 0.9 [0.04]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
4ZO2	C2	0.0 ± 0.0	0.0 ± 0.0 [0.00]	1.3 ± 1.1 [0.14]	2.4 ± 1.3 [0.94]	19.0 ± 20.2	0.0 ± 0.0 [0.00]	1.9 ± 1.2 [0.31]	2.0 ± 1.4 [0.87]
1SG4	C3	48.1 ± 5.2	5.0 ± 0.0 [1.00]	5.0 ± 0.0 [1.00]	10.0 ± 0.2 [1.00]	19.7 ± 6.0	0.0 ± 0.1 [0.00]	1.6 ± 1.1 [0.20]	3.4 ± 1.5 [0.98]
4CO0	C3	0.0 ± 0.0	0.0 ± 0.0 [0.00]	5.0 ± 0.0 [1.00]	10.0 ± 0.0 [1.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	5.0 ± 0.2 [1.00]	9.4 ± 0.8 [1.00]
4D7Y	C3	32.9 ± 2.8	5.0 ± 0.0 [1.00]	5.0 ± 0.0 [1.00]	10.0 ± 0.3 [1.00]	14.6 ± 2.2	1.4 ± 1.1 [0.17]	3.4 ± 1.1 [0.82]	7.5 ± 1.4 [1.00]
4XLA	C3	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
5I6N	C3	77.4 ± 11.2	4.9 ± 0.3 [1.00]	4.9 ± 0.3 [1.00]	8.6 ± 1.6 [1.00]	27.9 ± 11.1	0.0 ± 0.0 [0.00]	1.9 ± 1.1 [0.28]	3.5 ± 1.6 [0.99]
1P5B	C4	85.8 ± 7.7	5.0 ± 0.0 [1.00]	5.0 ± 0.0 [1.00]	10.0 ± 0.0 [1.00]	22.2 ± 4.6	1.1 ± 0.9 [0.07]	0.6 ± 0.8 [0.02]	1.2 ± 1.0 [0.70]
2O6N	C4	12.1 ± 0.7	5.0 ± 0.0 [1.00]	5.0 ± 0.0 [1.00]	10.0 ± 0.0 [1.00]	14.3 ± 2.3	0.8 ± 0.8 [0.04]	5.0 ± 0.0 [1.00]	10.0 ± 0.0 [1.00]
2V9N	C4	41.5 ± 3.9	5.0 ± 0.0 [1.00]	5.0 ± 0.0 [1.00]	10.0 ± 0.0 [1.00]	0.8 ± 2.5	0.0 ± 0.0 [0.00]	1.0 ± 1.0 [0.09]	1.0 ± 1.0 [0.64]

220

		Rosetta SymDock2				Rosetta SymDock			
PDBID	Symmetry	$\langle E_{1\%} \rangle$	Coarse-grained $\langle N5 \rangle$ [P($\langle N5 \rangle \geq 3$)]	Full protocol $\langle N5 \rangle$ [P($\langle N5 \rangle \geq 3$)]	Full protocol $\langle N10 \rangle$ [P($\langle N10 \rangle \geq 1$)]	$\langle E_{1\%} \rangle$	Coarse-grained $\langle N5 \rangle$ [P($\langle N5 \rangle \geq 3$)]	Full protocol $\langle N5 \rangle$ [P($\langle N5 \rangle \geq 3$)]	Full protocol $\langle N10 \rangle$ [P($\langle N10 \rangle \geq 1$)]
3V9O	C4	86.0 ± 4.8	5.0 ± 0.0 [1.00]	5.0 ± 0.0 [1.00]	10.0 ± 0.1 [1.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
4Z0G	C4	97.9 ± 14.5	3.6 ± 1.4 [0.77]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
1XB9	C5	93.0 ± 7.9	5.0 ± 0.0 [1.00]	5.0 ± 0.0 [1.00]	10.0 ± 0.0 [1.00]	14.9 ± 3.7	0.2 ± 0.5 [0.01]	0.4 ± 0.7 [0.02]	1.8 ± 1.3 [0.82]
4AVS	C5	38.5 ± 4.4	2.9 ± 1.2 [0.64]	5.0 ± 0.1 [1.00]	9.9 ± 0.4 [1.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	1.2 ± 1.0 [0.12]	2.6 ± 1.5 [0.94]
4U62	C5	100.0 ± 0.0	5.0 ± 0.1 [1.00]	5.0 ± 0.1 [1.00]	9.5 ± 1.1 [1.00]	34.9 ± 39.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
5A12	C5	99.9 ± 1.1	4.3 ± 0.8 [0.96]	5.0 ± 0.0 [1.00]	10.0 ± 0.2 [1.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.3 ± 0.6 [0.01]	0.9 ± 1.0 [0.59]
5LZH	C5	34.2 ± 2.8	5.0 ± 0.0 [1.00]	5.0 ± 0.0 [1.00]	10.0 ± 0.0 [1.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
1NLF	C6	95.4 ± 21.2	0.4 ± 0.7 [0.02]	4.8 ± 0.6 [0.98]	7.7 ± 1.7 [1.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	1.9 ± 1.4 [0.32]	2.0 ± 1.4 [0.87]
2XF7	C6	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.1 ± 0.3 [0.00]	0.7 ± 0.9 [0.46]	1.3 ± 1.5	0.0 ± 0.0 [0.00]	1.0 ± 0.9 [0.06]	1.6 ± 1.1 [0.82]
3H47	C6	6.1 ± 7.5	0.0 ± 0.0 [0.00]	4.6 ± 0.7 [0.99]	8.3 ± 1.4 [1.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	2.1 ± 1.2 [0.35]	3.6 ± 1.5 [0.99]
4OX6	C6	19.2 ± 1.3	5.0 ± 0.0 [1.00]	5.0 ± 0.0 [1.00]	10.0 ± 0.0 [1.00]	13.5 ± 0.8	5.0 ± 0.0 [1.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
4W64	C6	0.0 ± 0.0	0.0 ± 0.0 [0.00]	1.4 ± 1.1 [0.17]	4.1 ± 1.6 [0.99]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
1H64	C7	30.9 ± 2.4	5.0 ± 0.0 [1.00]	5.0 ± 0.0 [1.00]	10.0 ± 0.0 [1.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.7 ± 0.8 [0.03]	1.0 ± 1.0 [0.65]

221

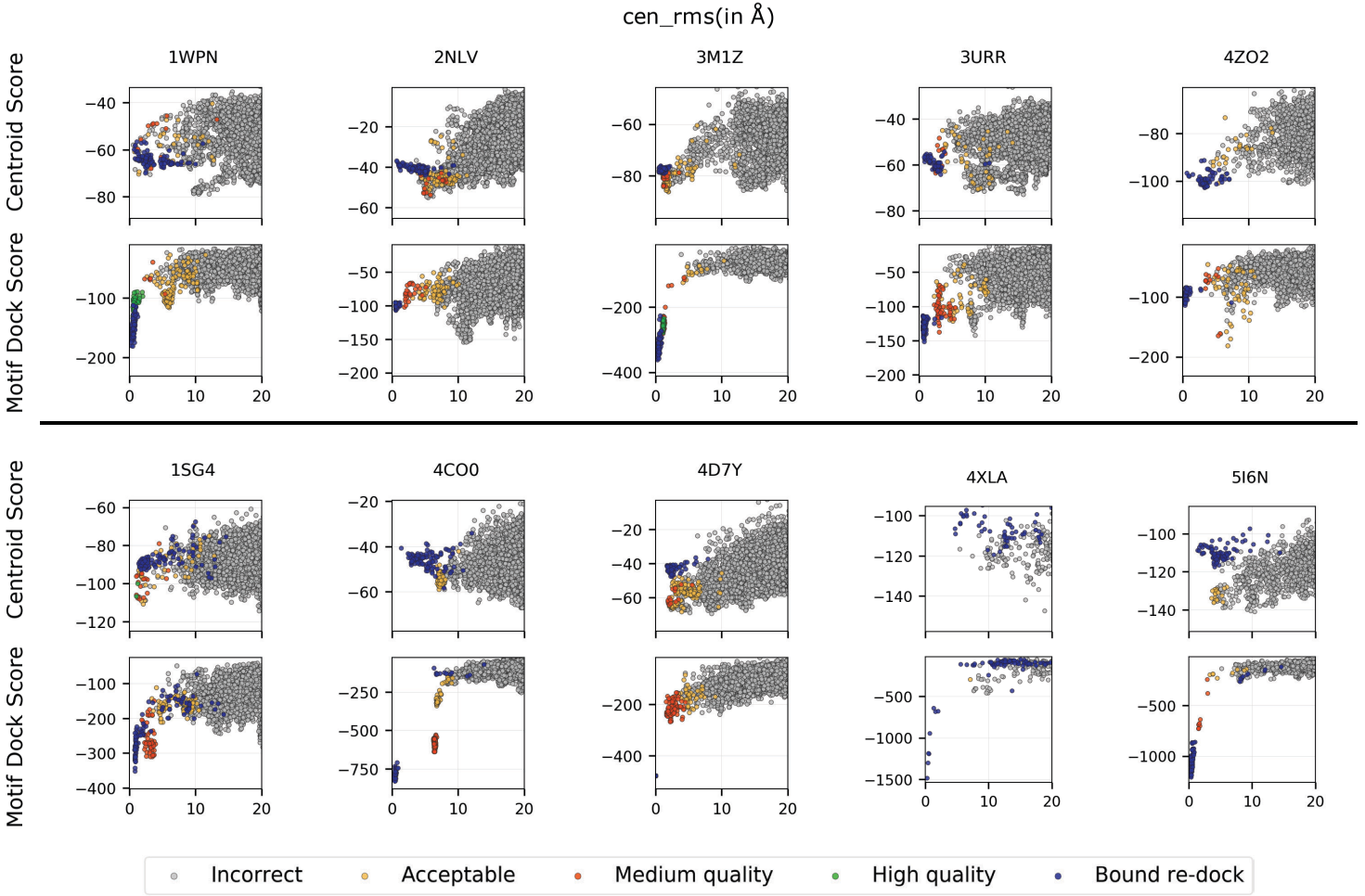
		Rosetta SymDock2				Rosetta SymDock			
PDBID	Symmetry	$\langle E_{1\%} \rangle$	Coarse-grained $\langle N5 \rangle$ [P($\langle N5 \rangle \geq 3$)]	Full protocol $\langle N5 \rangle$ [P($\langle N5 \rangle \geq 3$)]	Full protocol $\langle N10 \rangle$ [P($\langle N10 \rangle \geq 1$)]	$\langle E_{1\%} \rangle$	Coarse-grained $\langle N5 \rangle$ [P($\langle N5 \rangle \geq 3$)]	Full protocol $\langle N5 \rangle$ [P($\langle N5 \rangle \geq 3$)]	Full protocol $\langle N10 \rangle$ [P($\langle N10 \rangle \geq 1$)]
4OWK	C7	0.0 ± 0.1	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
3B8O	C8	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
4F87	C8	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
3P9A	C9	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
3ZQO	C9	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
1ORR	D2	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.2 ± 0.5 [0.12]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
1ZJZ	D2	62.0 ± 48.9	0.0 ± 0.0 [0.00]	5.0 ± 0.0 [1.00]	10.0 ± 0.0 [1.00]	53.3 ± 26.9	0.9 ± 0.9 [0.07]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
2BV4	D2	50.5 ± 23.7	0.8 ± 0.9 [0.05]	4.9 ± 0.4 [1.00]	9.1 ± 1.0 [1.00]	32.0 ± 31.6	0.0 ± 0.0 [0.00]	1.6 ± 1.1 [0.22]	2.2 ± 1.4 [0.91]
2VQR	D2	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
4OQC	D2	0.0 ± 0.0	0.0 ± 0.0 [0.00]	5.0 ± 0.0 [1.00]	10.0 ± 0.0 [1.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
1GXU	D3	0.0 ± 0.5	0.0 ± 0.0 [0.00]	0.0 ± 0.1 [0.00]	0.0 ± 0.1 [0.01]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
2BHQ	D3	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
2J5G	D3	85.7 ± 35.3	1.9 ± 1.3 [0.31]	3.2 ± 1.3 [0.70]	3.9 ± 1.9 [0.98]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]

222

		Rosetta SymDock2				Rosetta SymDock			
PDBID	Symm-etry	$\langle E_{1\%} \rangle$	Coarse-grained $\langle N5 \rangle$ [P($\langle N5 \rangle \geq 3$)]	Full protocol $\langle N5 \rangle$ [P($\langle N5 \rangle \geq 3$)]	Full protocol $\langle N10 \rangle$ [P($\langle N10 \rangle \geq 1$)]	$\langle E_{1\%} \rangle$	Coarse-grained $\langle N5 \rangle$ [P($\langle N5 \rangle \geq 3$)]	Full protocol $\langle N5 \rangle$ [P($\langle N5 \rangle \geq 3$)]	Full protocol $\langle N10 \rangle$ [P($\langle N10 \rangle \geq 1$)]
3QNS	D3	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
3V4F	D3	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]
2R8E	D4	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.5 ± 0.8 [0.03]	1.8 ± 1.3 [0.83]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.4 ± 0.6 [0.01]	1.2 ± 1.1 [0.68]
3R1M	D4	0.0 ± 0.0	0.0 ± 0.0 [0.00]	3.6 ± 1.4 [0.76]	4.0 ± 2.0 [0.99]	0.0 ± 0.0	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]	0.0 ± 0.0 [0.00]

5,000 decoys were generated by each protocol for each target. Bootstrapped *N5* and *N10* values (plus standard deviations), both after the coarse-grained phase and after the full protocol, are listed for each target. Bootstrapped enrichment values are also shown. Cases where bootstrapping gave ≥ 50% chance of success are shown in bold; success is defined as $\langle N5 \rangle \geq 3$, and $\langle N10 \rangle \geq 1$.

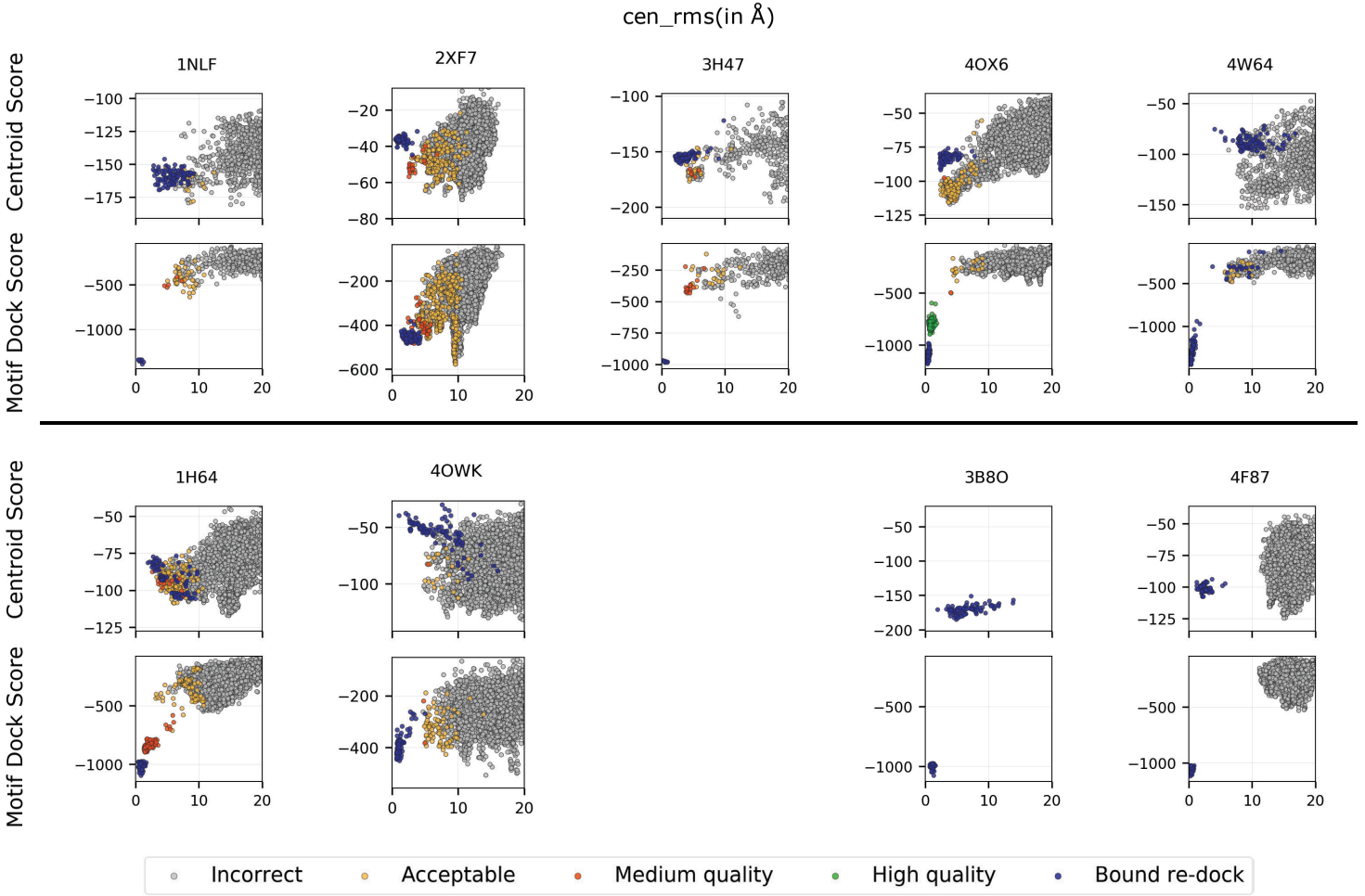
223


cen_rms(in Å)
1WPN
2NLV
3M1Z
3URR
4ZO2
1SG4
4CO0
4D7Y
4XLA
5I6N
Centroid Score
Motif Dock Score
Incorrect
Acceptable
Medium quality
High quality
Bound re-dock

224

cen_rms(in Å)
1P5B
2O6N
2V9N
3V9O
4Z0G
1XB9
4AVS
4U62
5A12
5LZH
Centroid Score
Motif Dock Score
Incorrect
Acceptable
Medium quality
High quality
Bound re-dock

225


cen_rms(in Å)
1NLF
2XF7
3H47
4OX6
4W64
1H64
4OWK
3B8O
4F87
Centroid Score
Motif Dock Score
Incorrect
Acceptable
Medium quality
High quality
Bound re-dock

226

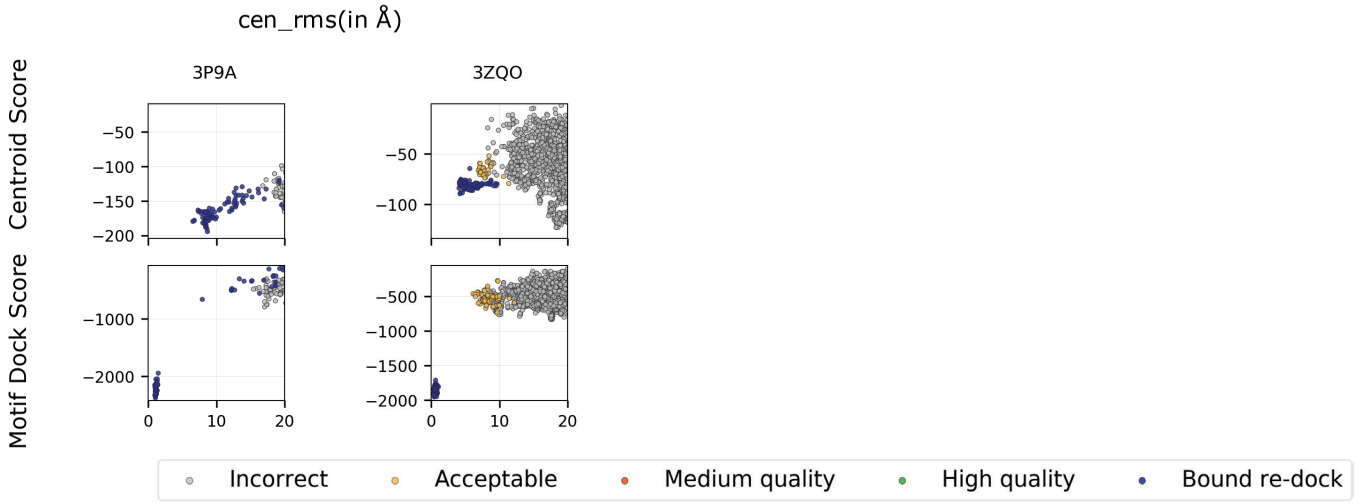


Figure B.1: Score versus RMSD plots after the coarse-grained phase for centroid score versus motif dock score for cyclic complexes during global docking.

cen_rms(in Å)
1ORR
1ZJZ
2BV4
2VQR
4OQC
1GXU
2BHQ
2J5G
3QNS
3V4F
Centroid Score
Motif Dock Score
Incorrect
Acceptable
Medium quality
High quality
Bound re-dock

228

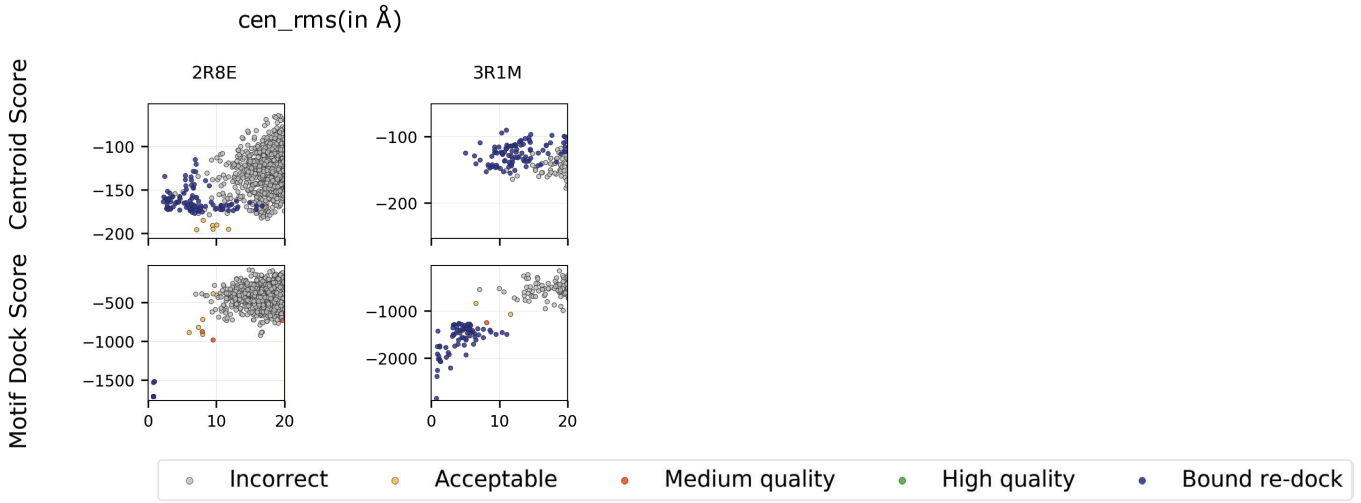
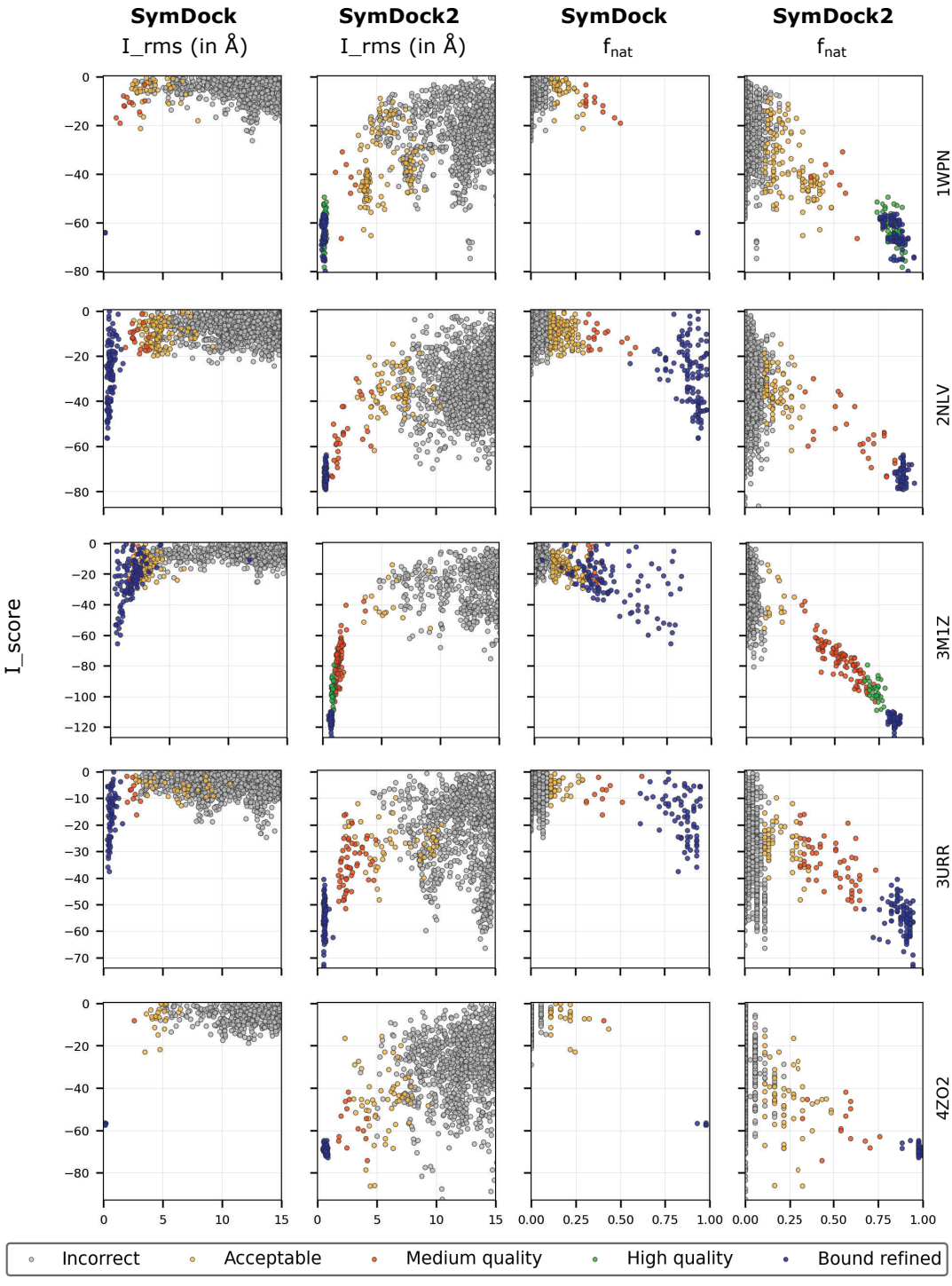
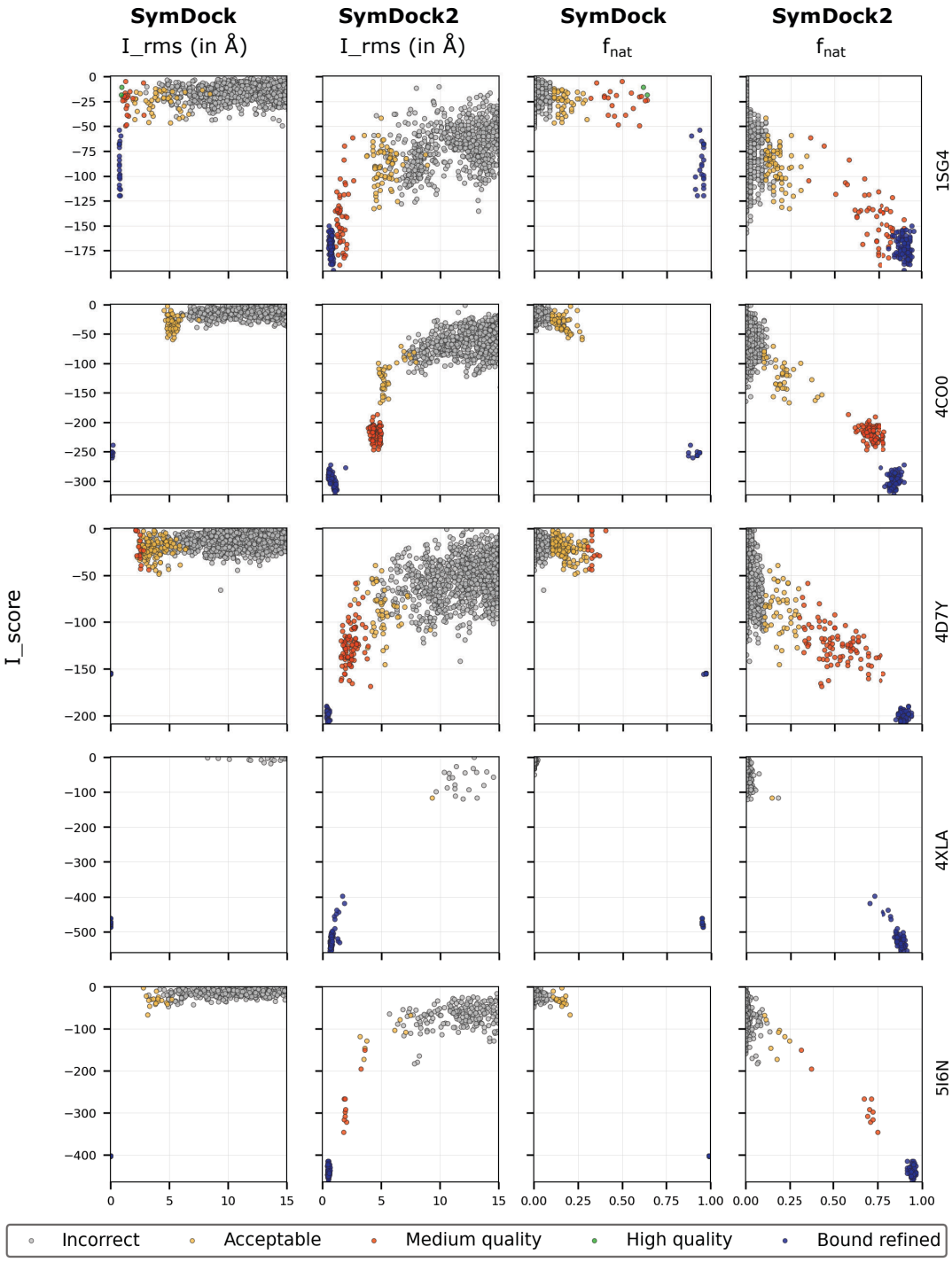
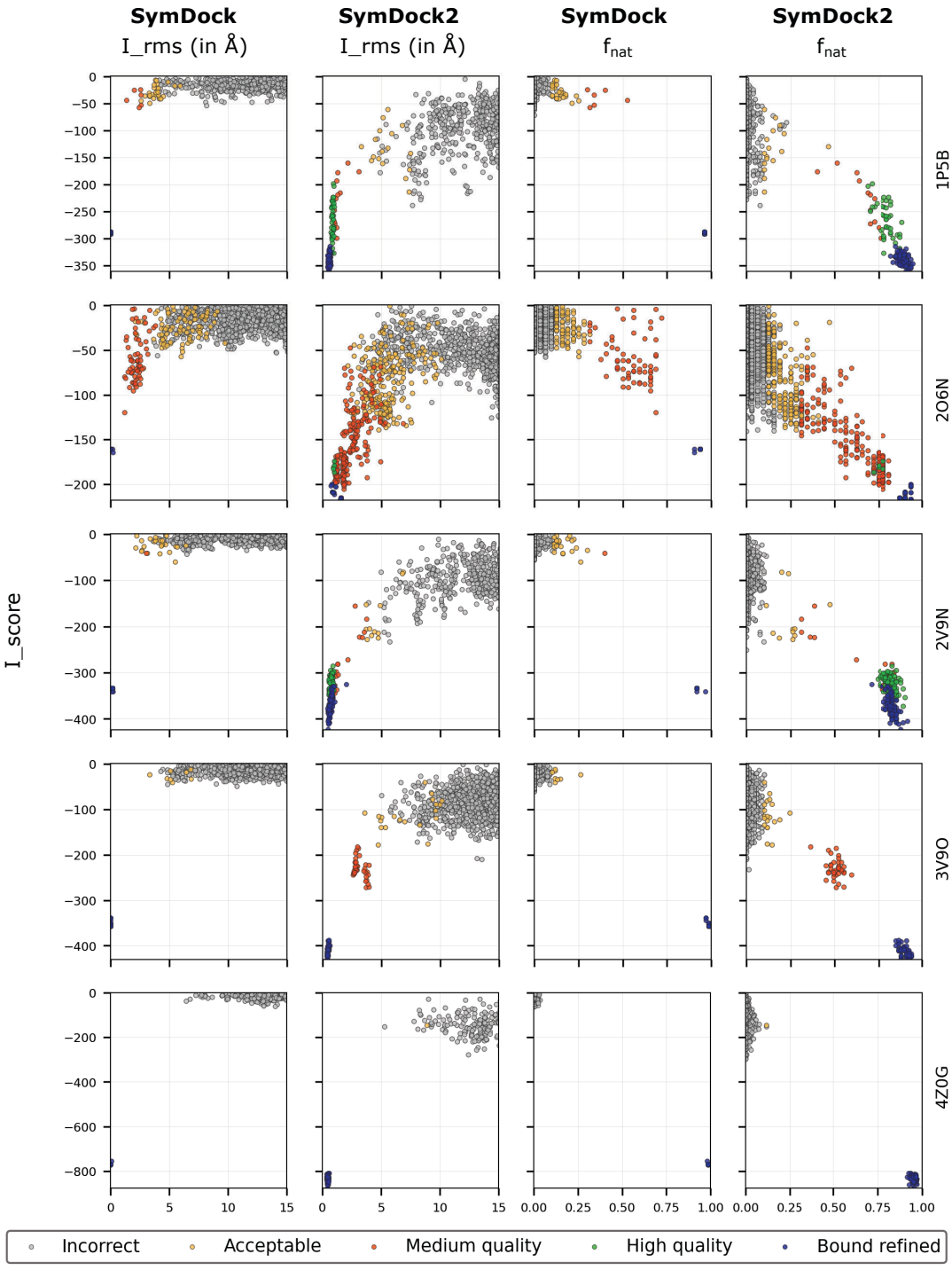
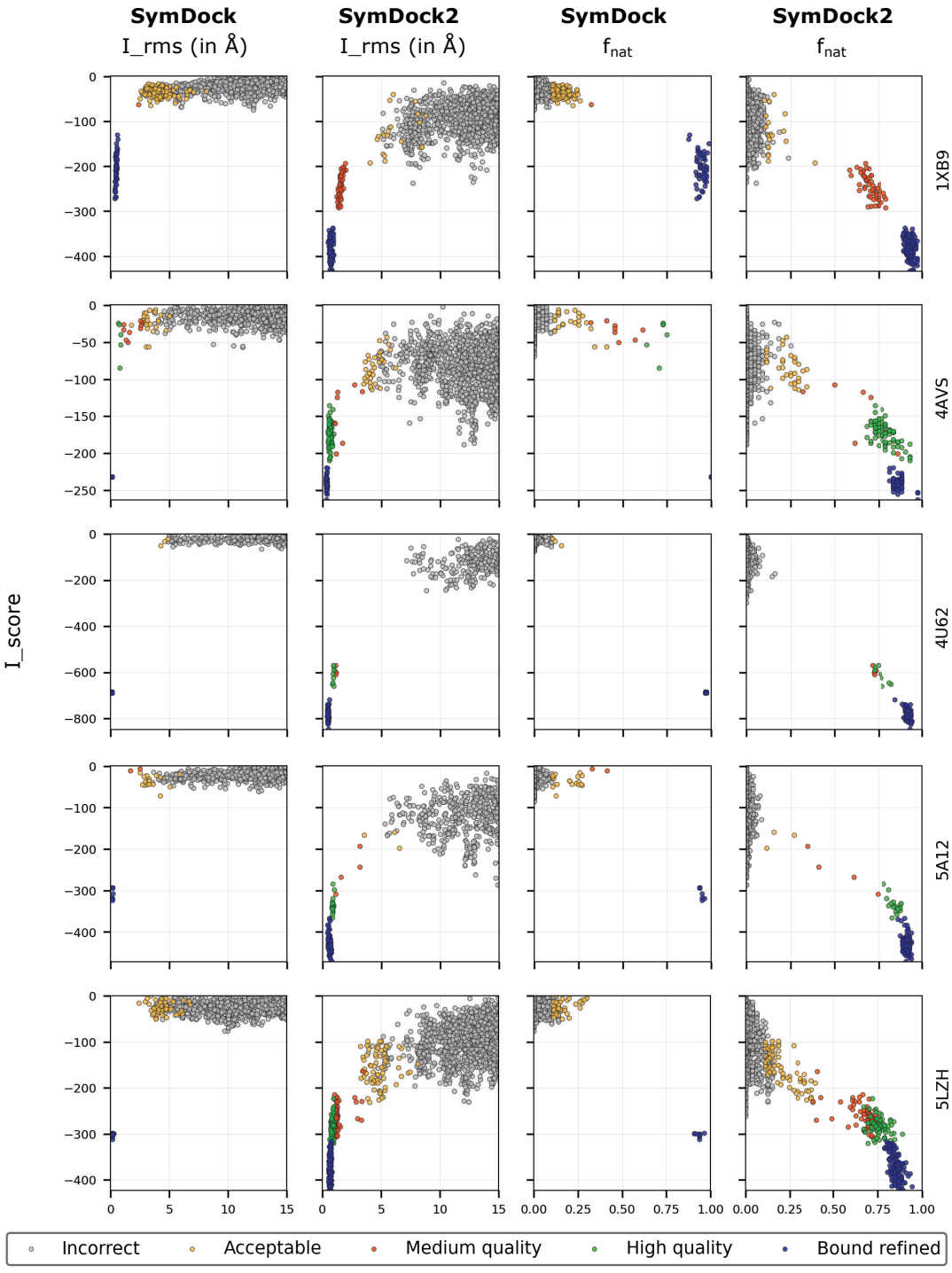


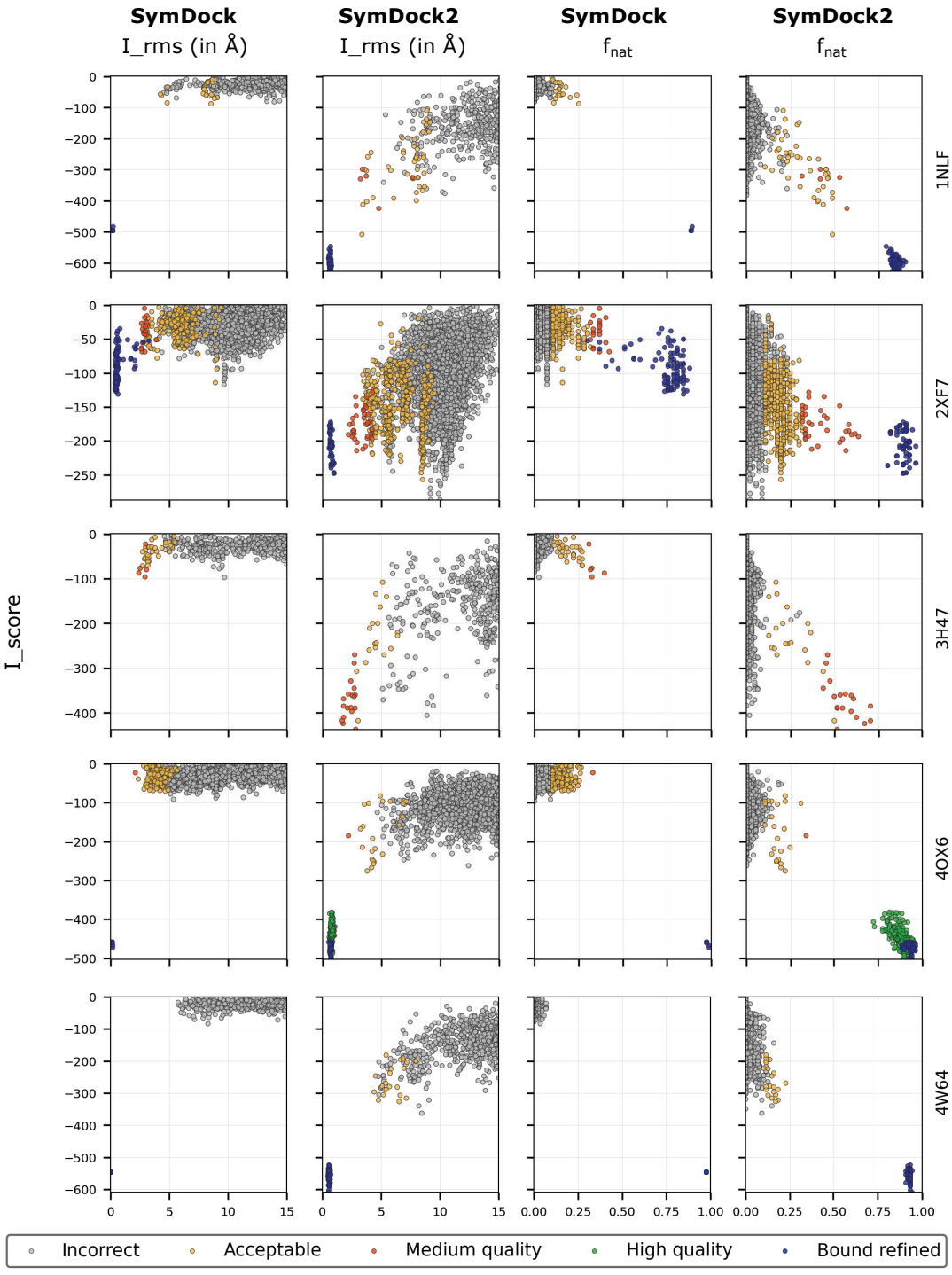
Figure B.2: Score versus RMSD plots after the coarse-grained phase for centroid score versus motif dock score for dihedral complexes during global docking.

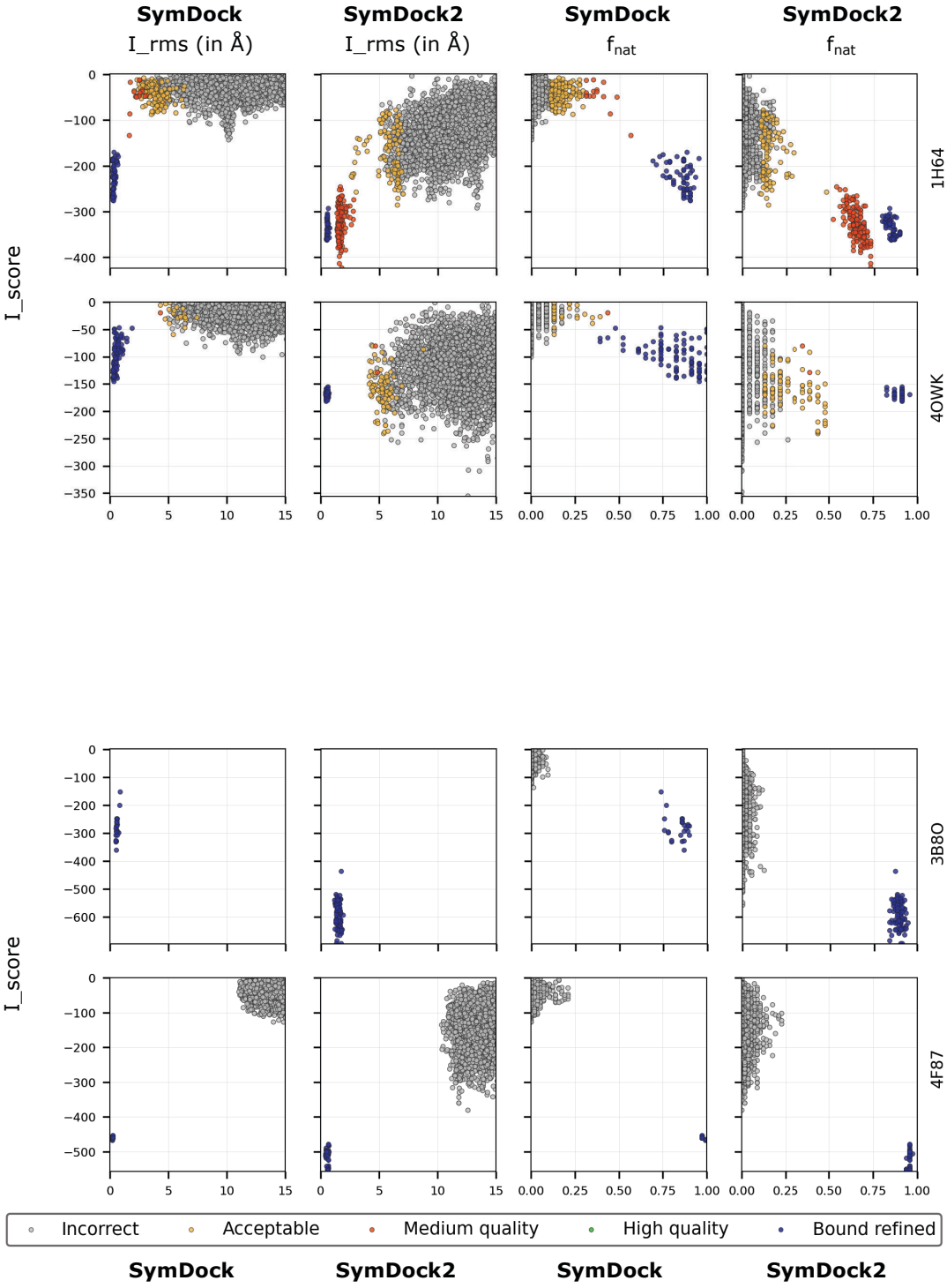

SymDock
I_rms (in Å)
SymDock2
I_rms (in Å)
SymDock
f_nat
SymDock2
f_nat
I_score
1WPN
2NLV
3M1Z
3URR
4ZO2
Incorrect
Acceptable
Medium quality
High quality
Bound refined


SymDock
I_rms (in Å)
SymDock2
I_rms (in Å)
SymDock
f_nat
SymDock2
f_nat
I_score
1SG4
4CO0
4D7Y
4XLA
5I6N
Incorrect
Acceptable
Medium quality
High quality
Bound refined


SymDock
I_rms (in Å)
SymDock2
I_rms (in Å)
SymDock
f_nat
SymDock2
f_nat
I_score
1P5B
2O6N
2V9N
3V9O
4Z0G
Incorrect
Acceptable
Medium quality
High quality
Bound refined


SymDock
I_rms (in Å)
SymDock2
I_rms (in Å)
SymDock
f_nat
SymDock2
f_nat
I_score
1XB9
4AVS
4U62
5A12
5LZH
Incorrect
Acceptable
Medium quality
High quality
Bound refined


SymDock
I_rms (in Å)
SymDock2
I_rms (in Å)
SymDock
f_nat
SymDock2
f_nat
I_score
1NLF
2XF7
3H47
4OX6
4W64
Incorrect
Acceptable
Medium quality
High quality
Bound refined


SymDock
I_rms (in Å)
SymDock2
I_rms (in Å)
SymDock
f_nat
SymDock2
f_nat
I_score
1H64
4OWK
3B8O
4F87
Incorrect
Acceptable
Medium quality
High quality
Bound refined
SymDock
SymDock2
SymDock
SymDock2

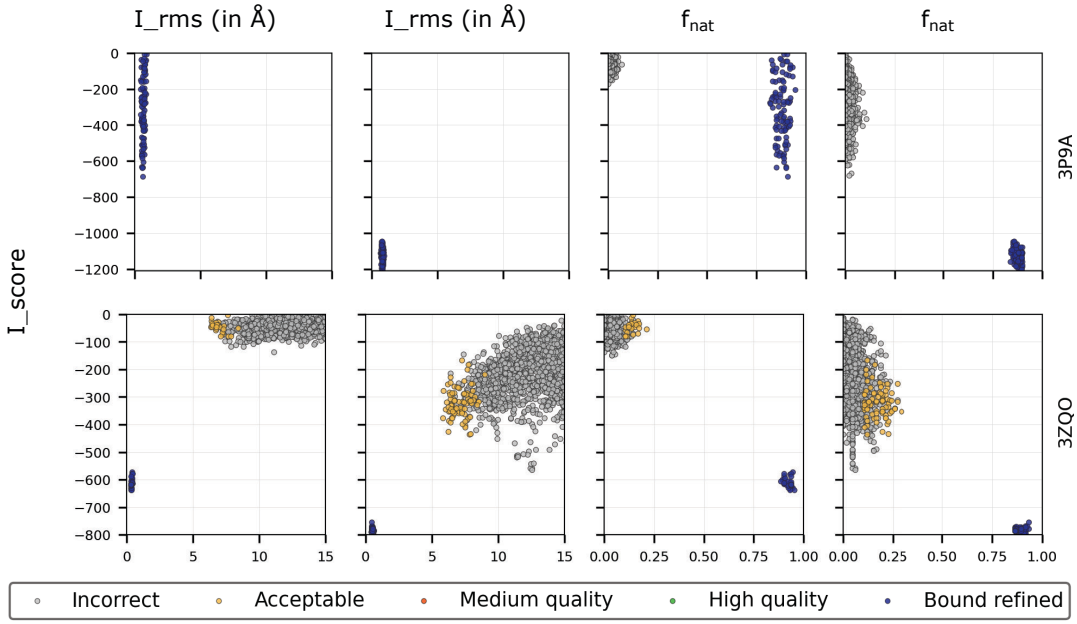
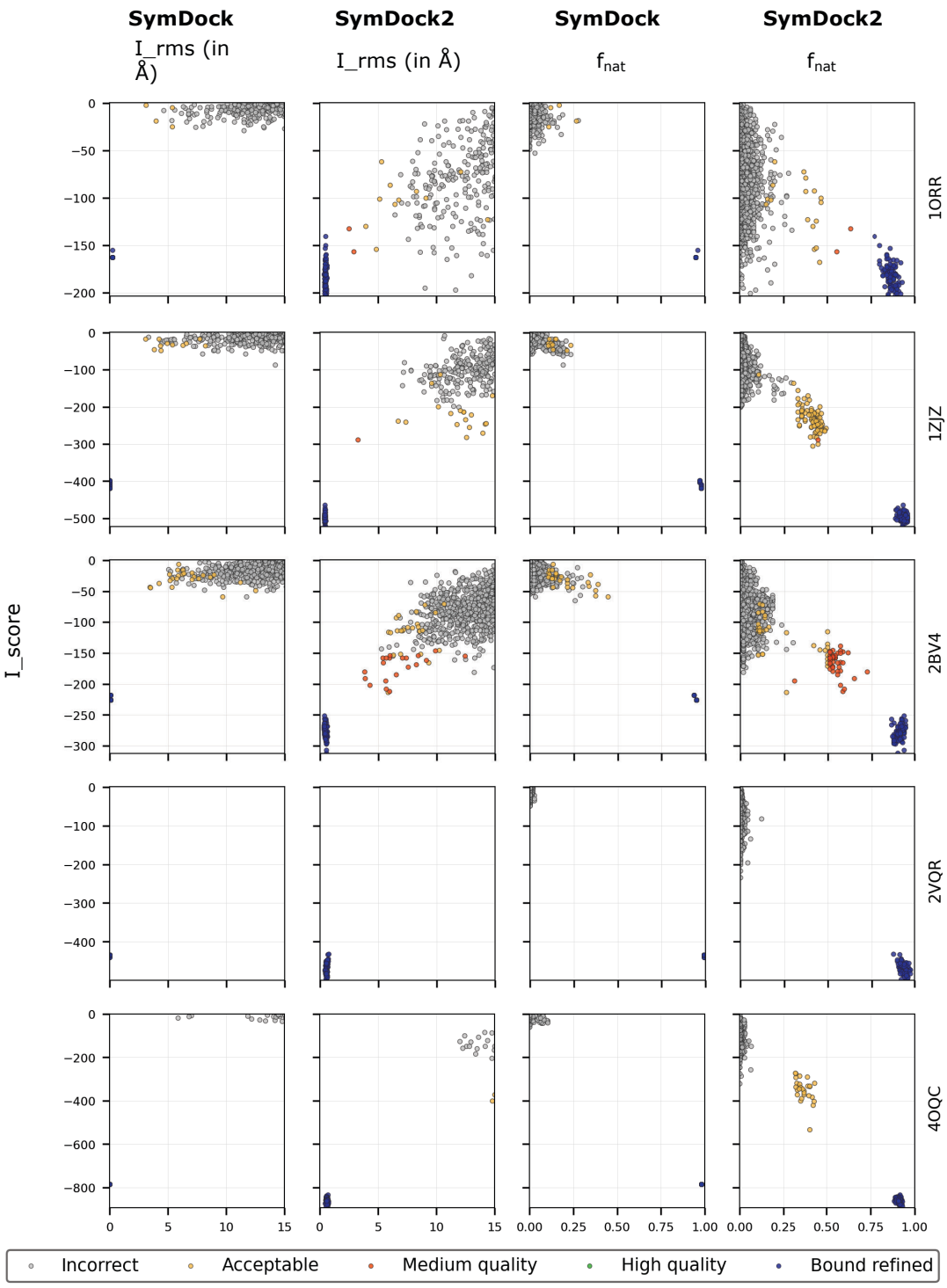
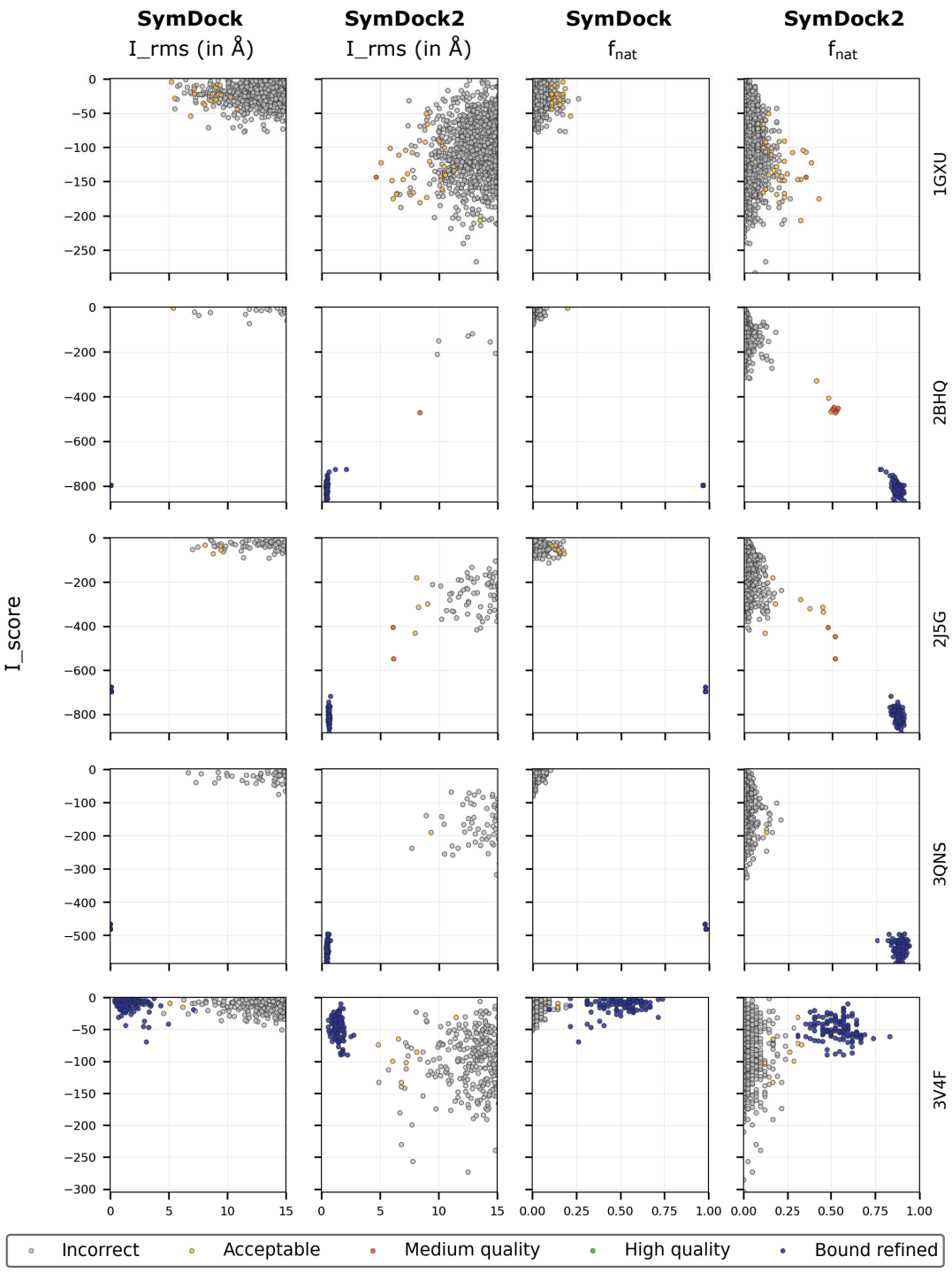


Figure B.3: Score versus RMSD plots & score versus f_{nat} plots after the full protocol for Rosetta SymDock and SymDock 2 for cyclic complexes for global docking.


SymDock
I_rms (in Å)
SymDock2
I_rms (in Å)
SymDock
f_nat
SymDock2
f_nat
I_score
1ORR
1ZJZ
2BV4
2VQR
4OQC
Incorrect
Acceptable
Medium quality
High quality
Bound refined


SymDock
I_rms (in Å)
SymDock2
I_rms (in Å)
SymDock
f_nat
SymDock2
f_nat
I_score
1GXU
2BHQ
2J5G
3QNS
3V4F
Incorrect
Acceptable
Medium quality
High quality
Bound refined

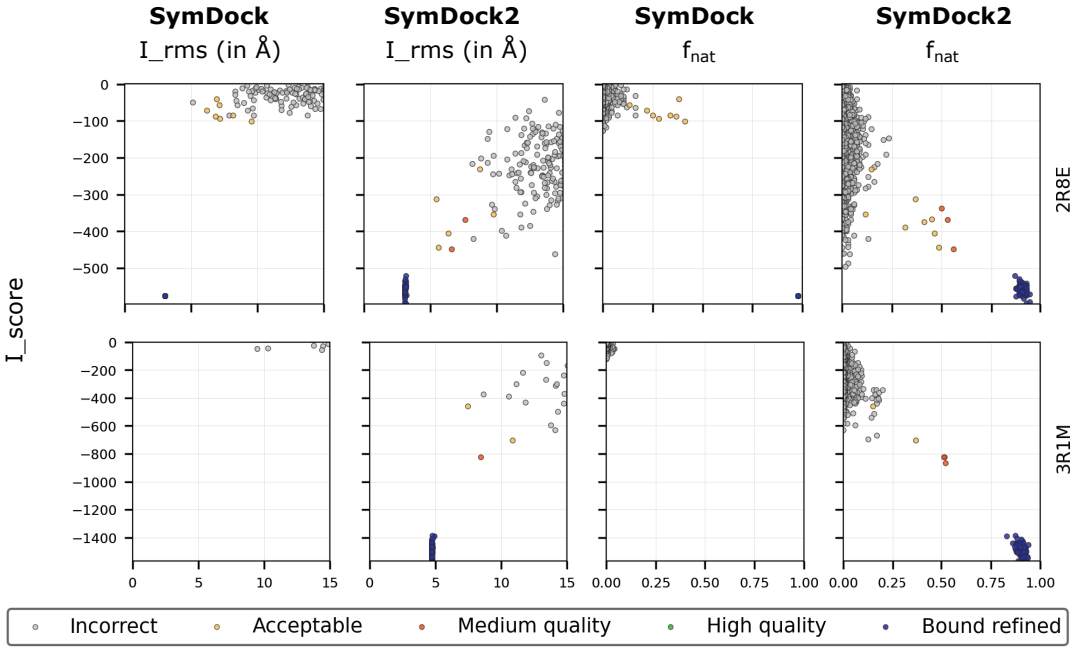


Figure B.4: Score versus RMSD plots & score versus f_{nat} plots after the full protocol for Rosetta SymDock and SymDock 2 for dihedral complexes for global docking.

Bibliography

1. Mosca, R., Pons, T., Céol, A., Valencia, A. & Aloy, P. Towards a detailed atlas of protein–protein interactions. *Curr. Opin. Struct. Biol.* **23,** 929–940 (2013).

2. Mosca, R., Céol, A. & Aloy, P. Interactome3D: adding structural details to protein networks. *Nat. Methods* **10,** 47–53 (2013).

3. Krawczyk, K., Kelm, S., Kovaltsuk, A., Galson, J. D., Kelly, D., Trück, J., Regep, C., Leem, J., Wong, W. K., Nowak, J., Snowden, J., Wright, M., Starkie, L., Scott-Tucker, A., Shi, J. & Deane, C. M. Structurally Mapping Antibody Repertoires. *Front. Immunol.* **9,** 1698 (2018).

4. Kilambi, K. P. & Gray, J. J. Structure-based cross-docking analysis of antibody–antigen interactions. *Sci. Rep.* **7,** 8145 (2017).

5. King, N. P., Sheffler, W., Sawaya, M. R., Vollmar, B. S., Sumida, J. P., Andre, I., Gonen, T., Yeates, T. O. & Baker, D. Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. *Science* **336,** 1171–1174 (2012).

6. Bale, J. B., Gonen, S., Liu, Y., Sheffler, W., Ellis, D., Thomas, C., Cascio, D., Yeates, T. O., Gonen, T., King, N. P. & Baker, D. Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **353,** 389–94 (2016).

7. Fallas, J. A., Ueda, G., Sheffler, W., Nguyen, V., McNamara, D. E., Sankaran, B., Pereira, J. H., Parmeggiani, F., Brunette, T. J., Cascio, D., Yeates, T. R., Zwart, P. & Baker, D. Computational design of self-assembling cyclic protein homo-oligomers. *Nat. Chem.* **9,** 353–360 (2017).

8. Nowak, R. P., DeAngelo, S. L., Buckley, D., He, Z., Donovan, K. A., An, J., Safaee, N., Jedrychowski, M. P., Ponthier, C. M., Ishoey, M., Zhang, T., Mancias, J. D., Gray, N. S., Bradner, J. E. & Fischer, E. S. Plasticity in binding confers selectivity in ligand-induced protein degradation. *Nat. Chem. Biol.* **14,** 706–714 (2018).

9. Greer, J. & Bush, B. L. Macromolecular shape and surface maps by solvent exclusion. *Proc. Natl. Acad. Sci.* **75,** 303–7 (1978).

10. Wodak, S. J. & Janin, J. Computer analysis of protein-protein interaction. *J. Mol. Biol.* **124,** 323–342 (1978).

11. Janin, J. & Wodak, S. J. Reaction pathway for the quaternary structure change in hemoglobin. *Biopolymers* **24,** 509–526 (1985).

12. Eaton, W. A., Henry, E. R. & Hofrichter, J. Application of linear free energy relations to protein conformational changes: the quaternary structural change of hemoglobin. *Proc. Natl. Acad. Sci.* **88,** 4472–5 (1991).

13. Rivetti, C., Mozzarelli, A., Rossi, G. L., Henry, E. R. & Eaton, W. A. Oxygen binding by single crystals of hemoglobin. *Biochemistry* **32,** 2888–2906 (1993).

14. Connolly, M. L. Shape complementarity at the hemoglobin α1β1 subunit interface.

Biopolymers **25,** 1229–1247 (1986).

15. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

16. Szilagyi, A. & Zhang, Y. Template-based structure modeling of protein–protein interactions. *Curr. Opin. Struct. Biol.* **24,** 10–23 (2014).

17. Aloy, P., Ceulemans, H., Stark, A. & Russell, R. B. The Relationship Between Sequence and Interaction Divergence in Proteins. *J. Mol. Biol.* **332,** 989–998 (2003).

18. Cukuroglu, E., Gursoy, A., Nussinov, R. & Keskin, O. Non-Redundant Unique Interface Structures as Templates for Modeling Protein Interactions. *PLoS One* **9,** e86738 (2014).

19. Anishchenko, I., Kundrotas, P. J., Tuzikov, A. V. & Vakser, I. A. Structural templates for comparative protein docking. *Proteins* **83,** 1563–1570 (2015).

20. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C. & Vakser, I. A. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci.* **89,** 2195–9 (1992).

21. Gabb, H. A., Jackson, R. M. & Sternberg, M. J. E. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* **272,** 106–120 (1997).

22. Tovchigrechko, A. & Vakser, I. A. GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res.* **34,** W310–W314 (2006).

23. Ravikant, D. V. S. & Elber, R. PIE—Efficient filters and coarse grained potentials for unbound protein–protein docking. *Proteins* **78,** 400–419 (2010).

24. Pierce, B. G., Hourai, Y. & Weng, Z. Accelerating Protein Docking in ZDOCK Using an Advanced 3D Convolution Library. *PLoS One* **6,** e24657 (2011).

25. Padhorny, D., Kazennov, A., Zerbe, B. S., Porter, K. A., Xia, B., Mottarella, S. E., Kholodov, Y., Ritchie, D. W., Vajda, S. & Kozakov, D. Protein-protein docking by fast generalized Fourier transforms on 5D rotational manifolds. *Proc. Natl. Acad. Sci.* **113,** E4286-93 (2016).

26. Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A. & Baker, D. Protein–Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations. *J. Mol. Biol.* **331,** 281–299 (2003).

27. Lorenzen, S. & Zhang, Y. Monte Carlo refinement of rigid-body protein docking structures with backbone displacement and side-chain optimization. *Protein Sci.* **16,** 2716–2725 (2007).

28. Zacharias, M. ATTRACT: Protein-protein docking in CAPRI using a reduced protein model. *Proteins* **60,** 252–256 (2005).

29. de Vries, S. J., van Dijk, M. & Bonvin, A. M. J. J. The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc.* **5,** 883–897 (2010).

30. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife* **3,**

e02030 (2014).

31. Hopf, T. A., Schärfe, C. P. I., Rodrigues, J. P. G. L. M., Green, A. G., Kohlbacher, O., Sander, C., Bonvin, A. M. J. J. & Marks, D. S. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **3,** (2014).

32. Xue, L. C., Rodrigues, J. P. G. L. M., Dobbs, D., Honavar, V. & Bonvin, A. M. J. J. Template-based protein–protein docking exploiting pairwise interfacial residue restraints. *Brief. Bioinform.* **18,** bbw027 (2016).

33. Jiménez-García, B., Pons, C., Svergun, D. I., Bernadó, P. & Fernández-Recio, J. pyDockSAXS: protein–protein complex structure by SAXS and computational docking. *Nucleic Acids Res.* **43,** W356–W361 (2015).

34. Knight, J. L., Mekler, V., Mukhopadhyay, J., Ebright, R. H. & Levy, R. M. Distance-Restrained Docking of Rifampicin and Rifamycin SV to RNA Polymerase Using Systematic FRET Measurements: Developing Benchmarks of Model Quality and Reliability. *Biophys. J.* **88,** 925–938 (2005).

35. Vreven, T., Schweppe, D. K., Chavez, J. D., Weisbrod, C. R., Shibata, S., Zheng, C., Bruce, J. E. & Weng, Z. Integrating Cross-Linking Experiments with Ab Initio Protein–Protein Docking. *J. Mol. Biol.* **430,** 1814–1828 (2018).

36. Roberts, V. A., Pique, M. E., Hsu, S. & Li, S. Combining H/D Exchange Mass Spectrometry and Computational Docking To Derive the Structure of Protein–Protein Complexes. *Biochemistry* **56,** 6329–6342 (2017).

37. Topf, M., Baker, M. L., John, B., Chiu, W. & Sali, A. Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy. *J. Struct. Biol.* **149,** 191–203 (2005).

38. Lasker, K., Sali, A. & Wolfson, H. J. Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. *Proteins* **78,** 3205–3211 (2010).

39. Janin, J., Henrick, K., Moult, J., Eyck, L. Ten, Sternberg, M. J. E., Vajda, S., Vakser, I. & Wodak, S. J. CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* **52,** 2–9 (2003).

40. Wodak, S. J. & Méndez, R. Prediction of protein-protein interactions: The CAPRI experiment, its evaluation and implications. *Curr. Opin. Struct. Biol.* **14,** 242–249 (2004).

41. Lensink, M. F. & Wodak, S. J. Docking and scoring protein interactions: CAPRI 2009. *Proteins* **78,** 3073–3084 (2010).

42. Lensink, M. F. & Wodak, S. J. Blind predictions of protein interfaces by docking calculations in CAPRI. *Proteins* **78,** 3085–3095 (2010).

43. Lensink, M. F. & Wodak, S. J. Docking, scoring, and affinity prediction in CAPRI. *Proteins* **81,** 2082–95 (2013).

44. Lensink, M. F., Moal, I. H., Bates, P. A., Kastritis, P. L., Melquiond, A. S. J., Karaca, E., Schmitz, C., van Dijk, M., Bonvin, A. M. J. J., Eisenstein, M., Jiménez-García, B., Grosdidier, S., Solernou, A., Pérez-Cano, L., Pallara, C., Fernández-Recio, J., Xu, J.,

Muthu, P., Praneeth Kilambi, K., Gray, J. J., Grudinin, S., Derevyanko, G., Mitchell, J. C., Wieting, J., Kanamori, E., Tsuchiya, Y., Murakami, Y., Sarmiento, J., Standley, D. M., Shirota, M., Kinoshita, K., Nakamura, H., Chavent, M., Ritchie, D. W., Park, H., Ko, J., Lee, H., Seok, C., Shen, Y., Kozakov, D., Vajda, S., Kundrotas, P. J., Vakser, I. A., Pierce, B. G., Hwang, H., Vreven, T., Weng, Z., Buch, I., Farkash, E., Wolfson, H. J., Zacharias, M., Qin, S., Zhou, H.-X., Huang, S.-Y., Zou, X., Wojdyla, J. A., Kleanthous, C. & Wodak, S. J. Blind prediction of interfacial water positions in CAPRI. *Proteins* **82,** 620–632 (2014).

45. Lensink, M. F., Velankar, S. & Wodak, S. J. Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. *Proteins* **85,** 359–377 (2017).

46. Papoian, G. A., Ulander, J. & Wolynes, P. G. Role of Water Mediated Interactions in Protein−Protein Recognition Landscapes. *J. Am. Chem. Soc.* **125,** 9170–9178 (2003).

47. Mukherjee, S., Nithin, C., Divakaruni, Y. & Bahadur, R. P. Dissecting water binding sites at protein–protein interfaces: a lesson from the atomic structures in the Protein Data Bank. *J. Biomol. Struct. Dyn.* 1–16 (2018). doi:10.1080/07391102.2018.1453379

48. Zacharias, M. Accounting for conformational changes during protein–protein docking. *Curr. Opin. Struct. Biol.* **20,** 180–186 (2010).

49. Estrin, M. & Wolfson, H. J. SnapDock—template-based docking by Geometric Hashing. *Bioinformatics* **33,** i30–i36 (2017).

50. Karaca, E., Melquiond, A. S. J., de Vries, S. J., Kastritis, P. L. & Bonvin, A. M. J. J.

Building macromolecular assemblies by information-driven docking: introducing the HADDOCK multibody docking server. *Mol. Cell. Proteom.* **9,** 1784–94 (2010).

51. Levy, E. D., Pereira-Leal, J. B., Chothia, C. & Teichmann, S. A. 3D Complex: A Structural Classification of Protein Complexes. *PLoS Comput. Biol.* **2,** e155 (2006).

52. Levinthal, C. How to Fold Graciously. in *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois* 22–24 (1969).

53. Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **77,** 778–795 (2009).

54. Gabdoulline, R. R. & Wade, R. C. Biomolecular diffusional association. *Curr. Opin. Struct. Biol.* **12,** 204–213 (2002).

55. Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der Dtsch. Chem. Gesellschaft* **27,** 2985–2993 (1894).

56. Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci.* **44,** 98–104 (1958).

57. Monod, J., Wyman, J. & Changeux, J.-P. On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12,** 88–118 (1965).

58. Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **35,** 539–546 (2010).

59. Wang, C., Schueler-Furman, O. & Baker, D. Improved side-chain modeling for protein-protein docking. *Protein Sci.* **14,** 1328–1339 (2005).

60. Mashiach, E., Nussinov, R. & Wolfson, H. J. FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking. *Nucleic Acids Res.* **38,** W457–W461 (2010).

61. Schindler, C. E. M., de Vries, S. J. & Zacharias, M. iATTRACT: Simultaneous global and local interface optimization for protein-protein docking refinement. *Proteins Struct. Funct. Bioinforma.* **83,** 248–258 (2015).

62. Kuroda, D. & Gray, J. J. Pushing the Backbone in Protein-Protein Docking. *Structure* **24,** 1821–1829 (2016).

63. Zacharias, M. Accounting for conformational changes during protein–protein docking. *Curr. Opin. Struct. Biol.* **20,** 180–186 (2010).

64. Goodsell, D. S. & Olson, A. J. Structural Symmetry and Protein Function. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 105–153 (2000).

65. Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H. J. PatchDock and SymmDock: Servers for rigid and symmetric docking. *Nucleic Acids Res.* **33,** 363–367 (2005).

66. Pierce, B., Tong, W. & Weng, Z. M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. *Bioinformatics* **21,** 1472–1478 (2005).

67. André, I., Bradley, P., Wang, C. & Baker, D. Prediction of the structure of symmetrical

protein assemblies. *Proc. Natl. Acad. Sci.* **104,** 17656–61 (2007).

68. Baek, M., Park, T., Heo, L., Park, C. & Seok, C. GalaxyHomomer: a web server for protein homo-oligomer structure prediction from a monomer sequence or structure. *Nucleic Acids Res.* **45,** W320–W324 (2017).

69. Yan, Y., Tao, H. & Huang, S.-Y. HSYMDOCK: a docking web server for predicting the structure of protein homo-oligomers with Cn or Dn symmetry. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gky398

70. Das, R., André, I., Shen, Y., Wu, Y., Lemak, A., Bansal, S., Arrowsmith, C. H., Szyperski, T. & Baker, D. Simultaneous prediction of protein folding and docking at high resolution. *Proc. Natl. Acad. Sci.* **106,** 18978–83 (2009).

71. Levy, E. D., Erba, E. B., Robinson, C. V. & Teichmann, S. A. Assembly reflects evolution of protein complexes. *Nature* **453,** 1262–1265 (2008).

72. Lensink, M. F., Velankar, S., Kryshtafovych, A., Huang, S.-Y., Schneidman-Duhovny, D., Sali, A., Segura, J., Fernandez-Fuentes, N., Viswanath, S., Elber, R., Grudinin, S., Popov, P., Neveu, E., Lee, H., Baek, M., Park, S., Heo, L., Rie Lee, G., Seok, C., Qin, S., Zhou, H.-X., Ritchie, D. W., Maigret, B., Devignes, M.-D., Ghoorah, A., Torchala, M., Chaleil, R. A. G., Bates, P. A., Ben-Zeev, E., Eisenstein, M., Negi, S. S., Weng, Z., Vreven, T., Pierce, B. G., Borrman, T. M., Yu, J., Ochsenbein, F., Guerois, R., Vangone, A., Rodrigues, J. P. G. L. M., van Zundert, G., Nellen, M., Xue, L., Karaca, E., Melquiond, A. S. J., Visscher, K., Kastritis, P. L., Bonvin, A. M. J. J., Xu, X., Qiu, L.,

Yan, C., Li, J., Ma, Z., Cheng, J., Zou, X., Shen, Y., Peterson, L. X., Kim, H.-R., Roy, A., Han, X., Esquivel-Rodriguez, J., Kihara, D., Yu, X., Bruce, N. J., Fuller, J. C., Wade, R. C., Anishchenko, I., Kundrotas, P. J., Vakser, I. A., Imai, K., Yamada, K., Oda, T., Nakamura, T., Tomii, K., Pallara, C., Romero-Durana, M., Jiménez-García, B., Moal, I. H., Férnandez-Recio, J., Joung, J. Y., Kim, J. Y., Joo, K., Lee, J., Kozakov, D., Vajda, S., Mottarella, S., Hall, D. R., Beglov, D., Mamonov, A., Xia, B., Bohnuud, T., Del Carpio, C. A., Ichiishi, E., Marze, N., Kuroda, D., Roy Burman, S. S., Gray, J. J., Chermak, E., Cavallo, L., Oliva, R., Tovchigrechko, A. & Wodak, S. J. Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins* **84,** 323–348 (2016).

73. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K. W., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y.-E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D. & Bradley, P. Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods Enzymol.* **487,** 545–574 (2011).

74. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181,** 223–30 (1973).

75. Li, Z. & Scheraga, H. A. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci.* **84,** 6611–5 (1987).

76. Zhang, Z., Lange, O. F., Baker, D., DiMaio, F. & Song, Y. Replica Exchange Improves Sampling in Low-Resolution Docking Stage of RosettaDock. *PLoS One* **8,** e72096 (2013).

77. Park, H., Bradley, P., Greisen, P., Liu, Y., Mulligan, V. K., Kim, D. E., Baker, D. & DiMaio, F. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **12,** 6201–6212 (2016).

78. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T. & Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13,** 3031–3048 (2017).

79. Marze, N. A., Roy Burman, S. S., Sheffler, W., Gray, J. J. & Valencia, A. Efficient flexible backbone protein–protein docking for challenging targets. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty355

80. Chu, X., Gan, L., Wang, E. & Wang, J. Quantifying the topography of the intrinsic energy landscape of flexible biomolecular recognition. *Proc. Natl. Acad. Sci.* **110,** E2342-51 (2013).

81. Vakser, I. A. Protein-protein docking: from interaction to interactome. *Biophys. J.* **107,** 1785–93 (2014).

82. Hwang, H., Vreven, T., Janin, J. & Weng, Z. Protein-protein docking benchmark version 4.0. *Proteins* **78,** 3111–3114 (2010).

83. Vreven, T., Moal, I. H., Vangone, A., Pierce, B. G., Kastritis, P. L., Torchala, M., Chaleil, R., Jiménez-García, B., Bates, P. A., Fernandez-Recio, J., Bonvin, A. M. J. J. & Weng, Z. Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J. Mol. Biol.* **427,** 3031–3041 (2015).

84. Janin, J. & Wodak, S. J. Computer analysis of protein-protein interaction. *J. Mol. Biol.* **124,** 323–342 (1978).

85. Venkatraman, V. & Ritchie, D. W. Flexible protein docking refinement using pose-dependent normal mode analysis. *Proteins* **80,** 2262–2274 (2012).

86. Moal, I. H. & Bates, P. A. SwarmDock and the Use of Normal Modes in Protein-Protein Docking. *Int. J. Mol. Sci.* **11,** 3623–3648 (2010).

87. Grünberg, R., Leckner, J. & Nilges, M. Complementarity of Structure Ensembles in Protein-Protein Binding. *Structure* **12,** 2125–2136 (2004).

88. Chaudhury, S. & Gray, J. J. Conformer Selection and Induced Fit in Flexible Backbone Protein-Protein Docking Using Computational and NMR Ensembles. *J. Mol. Biol.* **381,** 1068–1087 (2008).

89. Trellet, M., Melquiond, A. S. J., Bonvin, A. M. J. J., Wodak, S. & Bhat, T. A Unified Conformational Selection and Induced Fit Approach to Protein-Peptide Docking.

PLoS One **8,** e58769 (2013).

90. Zhang, Z., Ehmann, U. & Zacharias, M. Monte Carlo replica-exchange based ensemble docking of protein conformations. *Proteins* **85,** 924–937 (2017).

91. Lensink, M. F., Velankar, S., Kryshtafovych, A., Huang, S.-Y., Schneidman-Duhovny, D., Sali, A., Segura, J., Fernandez-Fuentes, N., Viswanath, S., Elber, R., Grudinin, S., Popov, P., Neveu, E., Lee, H., Baek, M., Park, S., Heo, L., Rie Lee, G., Seok, C., Qin, S., Zhou, H.-X., Ritchie, D. W., Maigret, B., Devignes, M.-D., Ghoorah, A., Torchala, M., Chaleil, R. A. G., Bates, P. A., Ben-Zeev, E., Eisenstein, M., Negi, S. S., Weng, Z., Vreven, T., Pierce, B. G., Borrman, T. M., Yu, J., Ochsenbein, F., Guerois, R., Vangone, A., Rodrigues, J. P. G. L. M., van Zundert, G., Nellen, M., Xue, L., Karaca, E., Melquiond, A. S. J., Visscher, K., Kastritis, P. L., Bonvin, A. M. J. J., Xu, X., Qiu, L., Yan, C., Li, J., Ma, Z., Cheng, J., Zou, X., Shen, Y., Peterson, L. X., Kim, H.-R., Roy, A., Han, X., Esquivel-Rodriguez, J., Kihara, D., Yu, X., Bruce, N. J., Fuller, J. C., Wade, R. C., Anishchenko, I., Kundrotas, P. J., Vakser, I. A., Imai, K., Yamada, K., Oda, T., Nakamura, T., Tomii, K., Pallara, C., Romero-Durana, M., Jiménez-García, B., Moal, I. H., Férnandez-Recio, J., Joung, J. Y., Kim, J. Y., Joo, K., Lee, J., Kozakov, D., Vajda, S., Mottarella, S., Hall, D. R., Beglov, D., Mamonov, A., Xia, B., Bohnuud, T., Del Carpio, C. A., Ichiishi, E., Marze, N., Kuroda, D., Roy Burman, S. S., Gray, J. J., Chermak, E., Cavallo, L., Oliva, R., Tovchigrechko, A. & Wodak, S. J. Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins* **84,** 323–348 (2016).

92. Lensink, M. F., Velankar, S. & Wodak, S. J. Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. *Proteins* **85,** 359–377 (2017).

93. Baaden, M. & Marrink, S. J. Coarse-grain modelling of protein-protein interactions. *Current Opinion in Structural Biology* **23,** 878–886 (Elsevier Current Trends, 2013).

94. Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E. & Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* **116,** 7898–7936 (2016).

95. Changeux, J.-P. & Edelstein, S. Conformational selection or induced fit? 50 years of debate resolved. *F1000 Biol. Rep.* **3,** 19 (2011).

96. Vogt, A. D. & Di Cera, E. Conformational Selection or Induced Fit? A Critical Appraisal of the Kinetic Mechanism. *Biochemistry* **51,** 5894–5902 (2012).

97. Xu, M. & Lill, M. A. Utilizing Experimental Data for Reducing Ensemble Size in Flexible-Protein Docking. *J. Chem. Inf. Model.* **52,** 187–198 (2012).

98. Pallara, C., Rueda, M., Abagyan, R. & Fernández-Recio, J. Conformational Heterogeneity of Unbound Proteins Enhances Recognition in Protein–Protein Encounters. *J. Chem. Theory Comput.* **12,** 3236–3249 (2016).

99. Gray, J. J., Moughon, S. E., Kortemme, T., Schueler-Furman, O., Misura, K. M. S., Morozov, A. V. & Baker, D. Protein-protein docking predictions for the CAPRI experiment. *Proteins* **52,** 118–122 (2003).

100. Daily, M. D., Masica, D., Sivasubramanian, A., Somarouthu, S. & Gray, J. J. CAPRI rounds 3-5 reveal promising successes and future challenges for RosettaDock. *Proteins*

60, 181–186 (2005).

101. Chaudhury, S., Sircar, A., Sivasubramanian, A., Berrondo, M. & Gray, J. J. Incorporating biochemical information and backbone flexibility in RosettaDock for CAPRI rounds 6-12. *Proteins* **69,** 793–800 (2007).

102. Sircar, A., Chaudhury, S., Kilambi, K. P., Berrondo, M. & Gray, J. J. A generalized approach to sampling backbone conformations with RosettaDock for CAPRI rounds 13-19. *Proteins* **78,** 3115–3123 (2010).

103. Kilambi, K. P., Pacella, M. S., Xu, J., Labonte, J. W., Porter, J. R., Muthu, P., Drew, K., Kuroda, D., Schueler-Furman, O., Bonneau, R. & Gray, J. J. Extending RosettaDock with water, sugar, and pH for prediction of complex structures and affinities for CAPRI rounds 20-27. *Proteins* **81,** 2201–2209 (2013).

104. Chaudhury, S., Berrondo, M., Weitzner, B. D., Muthu, P., Bergman, H. & Gray, J. J. Benchmarking and Analysis of Protein Docking Performance in Rosetta v3.2. *PLoS One* **6,** e22477 (2011).

105. Marze, N. A., Jeliazkov, J. R., Roy Burman, S. S., Boyken, S. E., DiMaio, F. & Gray, J. J. Modeling oblong proteins and water-mediated interfaces with RosettaDock in CAPRI rounds 28-35. *Proteins* **85,** 479–486 (2017).

106. DeBartolo, J., Dutta, S., Reich, L. & Keating, A. E. Predictive Bcl-2 family binding models rooted in experiment or structure. *J. Mol. Biol.* **422,** 124–44 (2012).

107. Tyka, M. D., Keedy, D. a., André, I., Dimaio, F., Song, Y., Richardson, D. C.,

Richardson, J. S. & Baker, D. Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **405,** 607–618 (2011).

108. Smith, C. A. & Kortemme, T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J. Mol. Biol.* **380,** 742–56 (2008).

109. Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O. & Bahar, I. Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys. J.* **80,** 505–515 (2001).

110. Vangone, A., Rodrigues, J. P. G. L. M., Xue, L. C., van Zundert, G. C. P., Geng, C., Kurkcuoglu, Z., Nellen, M., Narasimhan, S., Karaca, E., van Dijk, M., Melquiond, A. S. J., Visscher, K. M., Trellet, M., Kastritis, P. L. & Bonvin, A. M. J. J. Sense and simplicity in HADDOCK scoring: Lessons from CASP-CAPRI round 1. *Proteins* **85,** 417–423 (2017).

111. Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., Beglov, D. & Vajda, S. The ClusPro web server for protein–protein docking. *Nat. Protoc.* **12,** 255–278 (2017).

112. Pierce, B. G., Hourai, Y., Weng, Z., Vajda, S. & Jaroszewski, L. Accelerating Protein Docking in ZDOCK Using an Advanced 3D Convolution Library. *PLoS One* **6,** e24657 (2011).

113. Wodak, S. J. & Méndez, R. Prediction of protein–protein interactions: the CAPRI

experiment, its evaluation and implications. *Curr. Opin. Struct. Biol.* **14,** 242–249 (2004).

114. Greener, J. G., Filippis, I. & Sternberg, M. J. E. Predicting Protein Dynamics and Allostery Using Multi-Protein Atomic Distance Constraints. *Structure* **25,** 546–558 (2017).

115. Oliwa, T. & Shen, Y. cNMA: a framework of encounter complex-based normal mode analysis to model conformational changes in protein interactions. *Bioinformatics* **31,** i151–i160 (2015).

116. Wang, C., Bradley, P. & Baker, D. Protein–Protein Docking with Backbone Flexibility. *J. Mol. Biol.* **373,** 503–519 (2007).

117. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

118. Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E.-M., Khare, S. D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., Meiler, J. & Baker, D. RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. *PLoS One* **6,** e20161 (2011).

119. Méndez, R., Leplae, R., De Maria, L. & Wodak, S. J. Assessment of blind predictions of protein-protein interactions: Current status of docking methods. *Proteins* **52,** 51–67 (2003).

120. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical

assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins* **86,** 7–15 (2018).

121. Kilambi, K. P., Pacella, M. S., Xu, J., Labonte, J. W., Porter, J. R., Muthu, P., Drew, K., Kuroda, D., Schueler-Furman, O., Bonneau, R. & Gray, J. J. Extending RosettaDock with water, sugar, and pH for prediction of complex structures and affinities for CAPRI rounds 20-27. *Proteins* **81,** 2201–2209 (2013).

122. Marze, N. A., Jeliazkov, J. R., Roy Burman, S. S., Boyken, S. E., DiMaio, F. & Gray, J. J. Modeling oblong proteins and water-mediated interfaces with RosettaDock in CAPRI rounds 28-35. *Proteins* **85,** 479–486 (2017).

123. Lensink, M. F., Velankar, S., Baek, M., Heo, L., Seok, C. & Wodak, S. J. The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. *Proteins* **86,** 257–273 (2018).

124. Sircar, A. & Gray, J. J. SnugDock: Paratope Structural Optimization during Antibody-Antigen Docking Compensates for Errors in Antibody Homology Models. *PLoS Comput. Biol.* **6,** e1000644 (2010).

125. Singh, A. K., Menéndez-Conejero, R., San Martín, C. & van Raaij, M. J. Crystal Structure of the Fibre Head Domain of the Atadenovirus Snake Adenovirus 1. *PLoS One* **9,** e114373 (2014).

126. Nguyen, T. H., Ballmann, M. Z., Do, H. T., Truong, H. N., Benkő, M., Harrach, B. & van Raaij, M. J. Crystal structure of raptor adenovirus 1 fibre head and role of the beta-

hairpin in siadenovirus fibre head domains. *Virol. J.* **13,** 106 (2016).

127. Grötzinger, S. W., Karan, R., Strillinger, E., Bader, S., Frank, A., Al Rowaihi, I. S., Akal, A., Wackerow, W., Archer, J. A., Rueping, M., Weuster-Botz, D., Groll, M., Eppinger, J. & Arold, S. T. Identification and Experimental Characterization of an Extremophilic Brine Pool Alcohol Dehydrogenase from Single Amplified Genomes. *ACS Chem. Biol.* **13,** 161–170 (2018).

128. Kilambi, K. P., Reddy, K. & Gray, J. J. Protein-Protein Docking with Dynamic Residue Protonation States. *PLoS Comput. Biol.* **10,** e1004018 (2014).

129. Bule, P., Alves, V. D., Israeli-Ruimy, V., Carvalho, A. L., Ferreira, L. M. A., Smith, S. P., Gilbert, H. J., Najmudin, S., Bayer, E. A. & Fontes, C. M. G. A. Assembly of Ruminococcus flavefaciens cellulosome revealed by structures of two cohesin-dockerin complexes. *Sci. Rep.* **7,** 759 (2017).

130. Slutzki, M., Reshef, D., Barak, Y., Haimovitz, R., Rotem-Bamberger, S., Lamed, R., Bayer, E. A. & Schueler-Furman, O. Crucial roles of single residues in binding affinity, specificity, and promiscuity in the cellulosomal cohesin-dockerin interface. *J. Biol. Chem.* **290,** 13654–66 (2015).

131. Vignali, D. A. A. & Kuchroo, V. K. IL-12 family cytokines: immunological playmakers. *Nat. Immunol.* **13,** 722–728 (2012).

132. Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M., Pieper, U. & Sali, A. Comparative Protein Structure Modeling Using Modeller. *Curr.*

Protoc. Bioinforma. **15,** 5.6.1-5.6.30 (2006).

133. Song, Y., DiMaio, F., Wang, R. Y.-R., Kim, D., Miles, C., Brunette, T., Thompson, J. & Baker, D. High-Resolution Comparative Modeling with RosettaCM. *Structure* **21,** 1735–1742 (2013).

134. Schröder, J., Moll, J. M., Baran, P., Grötzinger, J., Scheller, J. & Floss, D. M. Non-canonical interleukin 23 receptor complex assembly: p40 protein recruits interleukin 12 receptor β1 via site II and induces p19/interleukin 23 receptor interaction via site III. *J. Biol. Chem.* **290,** 359–70 (2015).

135. Kryshtafovych, A., Albrecht, R., Baslé, A., Bule, P., Caputo, A. T., Carvalho, A. L., Chao, K. L., Diskin, R., Fidelis, K., Fontes, C. M. G. A., Fredslund, F., Gilbert, H. J., Goulding, C. W., Hartmann, M. D., Hayes, C. S., Herzberg, O., Hill, J. C., Joachimiak, A., Kohring, G.-W., Koning, R. I., Lo Leggio, L., Mangiagalli, M., Michalska, K., Moult, J., Najmudin, S., Nardini, M., Nardone, V., Ndeh, D., Nguyen, T.-H., Pintacuda, G., Postel, S., van Raaij, M. J., Roversi, P., Shimon, A., Singh, A. K., Sundberg, E. J., Tars, K., Zitzmann, N. & Schwede, T. Target highlights from the first post-PSI CASP experiment (CASP12, May-August 2016). *Proteins* **86,** 27–50 (2018).

136. Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., Beglov, D. & Vajda, S. The ClusPro web server for protein–protein docking. *Nat. Protoc.* **12,** 255–278 (2017).

137. Lori, C., Ozaki, S., Steiner, S., Böhm, R., Abel, S., Dubey, B. N., Schirmer, T., Hiller, S.

& Jenal, U. Cyclic di-GMP acts as a cell cycle oscillator to drive chromosome replication. *Nature* **523,** 236–239 (2015).

138. Culurgioni, S., Mari, S., Bonetti, P., Gallini, S., Bonetto, G., Brennich, M., Round, A., Nicassio, F. & Mapelli, M. Insc:LGN tetramers promote asymmetric divisions of mammary stem cells. *Nat. Commun.* **9,** 1025 (2018).

139. Leone, P., Roche, J., Vincent, M. S., Tran, Q. H., Desmyter, A., Cascales, E., Kellenberger, C., Cambillau, C. & Roussel, A. Type IX secretion system PorM and gliding machinery GldM form arches spanning the periplasmic space. *Nat. Commun.* **9,** 429 (2018).

140. Pardon, E., Laeremans, T., Triest, S., Rasmussen, S. G. F., Wohlkönig, A., Ruf, A., Muyldermans, S., Hol, W. G. J., Kobilka, B. K. & Steyaert, J. A general protocol for the generation of Nanobodies for structural biology. *Nat. Protoc.* **9,** 674–693 (2014).

141. Sircar, A., Sanni, K. A., Shi, J. & Gray, J. J. Analysis and Modeling of the Variable Region of Camelid Single-Domain Antibodies. *J. Immunol.* **186,** 6357–6367 (2011).

142. Weitzner, B. D., Jeliazkov, J. R., Lyskov, S., Marze, N., Kuroda, D., Frick, R., Adolf-Bryfogle, J., Biswas, N., Dunbrack, R. L. & Gray, J. J. Modeling and docking of antibody structures with Rosetta. *Nat. Protoc.* **12,** 401–416 (2017).

143. Moonens, K., Hamway, Y., Neddermann, M., Reschke, M., Tegtmeyer, N., Kruse, T., Kammerer, R., Mejías-Luque, R., Singer, B. B., Backert, S., Gerhard, M. & Remaut, H. Helicobacter pylori adhesin HopQ disrupts trans dimerization in human CEACAMs.

EMBO J. **37,** e98665 (2018).

144. Fedarovich, A., Tomberg, J., Nicholas, R. A. & Davies, C. Structure of the N-terminal domain of human CEACAM1: Binding target of the opacity proteins during invasion of Neisseria meningitidis and N. gonorrhoeae. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **62,** 971–979 (2006).

145. Huang, Y.-H., Zhu, C., Kondo, Y., Anderson, A. C., Gandhi, A., Russell, A., Dougan, S. K., Petersen, B.-S., Melum, E., Pertel, T., Clayton, K. L., Raab, M., Chen, Q., Beauchemin, N., Yazaki, P. J., Pyzik, M., Ostrowski, M. A., Glickman, J. N., Rudd, C. E., Ploegh, H. L., Franke, A., Petsko, G. A., Kuchroo, V. K. & Blumberg, R. S. CEACAM1 regulates TIM-3-mediated tolerance and exhaustion. *Nature* **517,** 386–390 (2015).

146. Javaheri, A., Kruse, T., Moonens, K., Mejías-Luque, R., Debraekeleer, A., Asche, C. I., Tegtmeyer, N., Kalali, B., Bach, N. C., Sieber, S. A., Hill, D. J., Königer, V., Hauck, C. R., Moskalenko, R., Haas, R., Busch, D. H., Klaile, E., Slevogt, H., Schmidt, A., Backert, S., Remaut, H., Singer, B. B. & Gerhard, M. Helicobacter pylori adhesin HopQ engages in a virulence-enhancing interaction with human CEACAMs. *Nat. Microbiol.* **2,** 16189 (2017).

147. Canutescu, A. A. & Dunbrack, R. L. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.* **12,** 963–972 (2003).

148. Huang, P.-S., Ban, Y.-E. A., Richter, F., Andre, I., Vernon, R., Schief, W. R. & Baker,

D. RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. *PLoS One* **6,** e24109 (2011).

149. Reignier, T., Oldenburg, J., Flanagan, M. L., Hamilton, G. A., Martin, V. K. & Cannon, P. M. Receptor use by the Whitewater Arroyo virus glycoprotein. *Virology* **371,** 439–446 (2008).

150. Shimon, A., Shani, O. & Diskin, R. Structural Basis for Receptor Selectivity by the Whitewater Arroyo Mammarenavirus. *J. Mol. Biol.* **429,** 2825–2839 (2017).

151. Bliven, S., Lafita, A., Parker, A., Capitani, G. & Duarte, J. M. Automated evaluation of quaternary structures from protein crystals. *PLoS Comput. Biol.* **14,** e1006104 (2018).

152. Skálová, T., Bláha, J., Harlos, K., Dušková, J., Koval', T., Stránský, J., Hašek, J., Vaněk, O. & Dohnálek, J. Four crystal structures of human LLT1, a ligand of human NKR-P1, in varied glycosylation and oligomerization states. *Acta Crystallogr. D. Biol. Crystallogr.* **71,** 578–91 (2015).

153. Shulami, S., Raz-Pasteur, A., Tabachnikov, O., Gilead-Gropper, S., Shner, I. & Shoham, Y. The L-Arabinan utilization system of Geobacillus stearothermophilus. *J. Bacteriol.* **193,** 2838–50 (2011).

154. Alhassid, A., Ben-David, A., Tabachnikov, O., Libster, D., Naveh, E., Zolotnitsky, G., Shoham, Y. & Shoham, G. Crystal structure of an inverting GH 43 1,5-alpha-L-arabinanase from Geobacillus stearothermophilus complexed with its substrate. *Biochem. J.* **422,** 73–82 (2009).

155. Vahedi-Faridi, A., Licht, A., Bulut, H., Scheffel, F., Keller, S., Wehmeier, U. F., Saenger, W. & Schneider, E. Crystal Structures of the Solute Receptor GacH of Streptomyces glaucescens in Complex with Acarbose and an Acarbose Homolog: Comparison with the Acarbose-Loaded Maltose-Binding Protein of Salmonella typhimurium. *J. Mol. Biol.* **397,** 709–723 (2010).

156. Labonte, J. W., Adolf-Bryfogle, J., Schief, W. R. & Gray, J. J. Residue-centric modeling and design of saccharide and glycoconjugate structures. *J. Comput. Chem.* **38,** 276–287 (2017).

157. Wojdyla, J. A., Fleishman, S. J., Baker, D. & Kleanthous, C. Structure of the Ultra-High-Affinity Colicin E2 DNase–Im2 Complex. *J. Mol. Biol.* **417,** 79–94 (2012).

158. Stein, A. & Kortemme, T. Improvements to Robotics-Inspired Conformational Sampling in Rosetta. *PLoS One* **8,** e63090 (2013).

159. Kandiah, E., Carriel, D., Perard, J., Malet, H., Bacia, M., Liu, K., Chan, S. W. S., Houry, W. A., Ollagnier de Choudens, S., Elsen, S. & Gutsche, I. Structural insights into the Escherichia coli lysine decarboxylases and molecular determinants of interaction with the AAA+ ATPase RavA. *Sci. Rep.* **6,** 24601 (2016).

160. Mosca, R., Pons, T., Céol, A., Valencia, A. & Aloy, P. Towards a detailed atlas of protein–protein interactions. *Curr. Opin. Struct. Biol.* **23,** 929–940 (2013).

161. Wodak, S. J., Vlasblom, J., Turinsky, A. L. & Pu, S. Protein–protein interaction networks: the puzzling riches. *Curr. Opin. Struct. Biol.* **23,** 941–953 (2013).

162. Vakser, I. A. Protein-protein docking: from interaction to interactome. *Biophys. J.* **107,** 1785–1793 (2014).

163. Wolynes, P. G. Symmetry and the energy landscapes of biomolecules. *Proc. Natl. Acad. Sci.* **93,** 14249–55 (1996).

164. Monod, J., Wyman, J. & Changeux, J.-P. On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12,** 88–118 (1965).

165. Perutz, M. F. Mechanisms of cooperativity and allosteric regulation in proteins. *Q. Rev. Biophys.* **22,** 139 (1989).

166. Garcia-Seisdedos, H., Empereur-Mot, C., Elad, N. & Levy, E. D. Proteins evolve on the edge of supramolecular self-assembly. *Nature* **548,** 244 (2017).

167. Crick, F. H. C. & Watson, J. D. in *The Nature of Viruses* **5,** 5–18 (Churchill London, 1957).

168. Chan, K.-Y., Gumbart, J., McGreevy, R., Watermeyer, J. M., Sewell, B. T. & Schulten, K. Symmetry-Restrained Flexible Fitting for Symmetric EM Maps. *Structure* **19,** 1211–1218 (2011).

169. Joseph, A. P., Malhotra, S., Burnley, T., Wood, C., Clare, D. K., Winn, M. & Topf, M. Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. *Methods* **100,** 42–49 (2016).

170. Yang, J. & Zhang, Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* **43,** W174–W181 (2015).

171. Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M., Pieper, U. & Sali, A. Comparative Protein Structure Modeling Using Modeller. *Curr. Protoc. Bioinforma.* **15,** 5.6.1-5.6.30 (2006).

172. DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D. & André, I. Modeling Symmetric Macromolecular Structures in Rosetta3. *PLoS One* **6,** e20450 (2011).

173. Kahraman, A., Herzog, F., Leitner, A., Rosenberger, G., Aebersold, R. & Malmström, L. Cross-Link Guided Molecular Modeling with ROSETTA. *PLoS One* **8,** e73411 (2013).

174. Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J. M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K. K., Lemak, A., Ignatchenko, A., Arrowsmith, C. H., Szyperski, T., Montelione, G. T., Baker, D. & Bax, A. Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl. Acad. Sci.* **105,** 4685–90 (2008).

175. Sønderby, P., Rinnan, Å., Madsen, J. J., Harris, P., Bukrinski, J. T. & Peters, G. H. J. Small-Angle X-ray Scattering Data in Combination with RosettaDock Improves the Docking Energy Landscape. *J. Chem. Inf. Model.* **57,** 2463–2475 (2017).

176. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife* **3,** e02030 (2014).

177. Zhang, Z., Lange, O. F., Baker, D., DiMaio, F. P. & Song, Y. Replica Exchange Improves Sampling in Low-Resolution Docking Stage of RosettaDock. *PLoS One* **8,**

e72096 (2013).

178. Méndez, R., Leplae, R., De Maria, L. & Wodak, S. J. Assessment of blind predictions of protein-protein interactions: Current status of docking methods. *Proteins* **52,** 51–67 (2003).

179. Go, N., Noguti, T. & Nishikawa, T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci.* **80,** 3696–3700 (1983).

180. Ritchie, D. W. & Grudinin, S. Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry. *J. Appl. Crystallogr.* **49,** 158–167 (2016).

181. Villar, G., Wilber, A. W., Williamson, A. J., Thiara, P., Doye, J. P. K., Louis, A. A., Jochum, M. N., Lewis, A. C. F. & Levy, E. D. Self-Assembly and Evolution of Homomeric Protein Complexes. *Phys. Rev. Lett.* **102,** 118106 (2009).

182. Park, H., Kim, D. E., Ovchinnikov, S., Baker, D. & DiMaio, F. Automatic structure prediction of oligomeric assemblies using Robetta in CASP12. *Proteins* **86,** 283–291 (2018).

183. Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L. & Schwede, T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci. Rep.* **7,** 10480 (2017).

184. Kryshtafovych, A., Albrecht, R., Baslé, A., Bule, P., Caputo, A. T., Carvalho, A. L., Chao, K. L., Diskin, R., Fidelis, K., Fontes, C. M. G. A., Fredslund, F., Gilbert, H. J., Goulding, C. W., Hartmann, M. D., Hayes, C. S., Herzberg, O., Hill, J. C., Joachimiak,

A., Kohring, G.-W., Koning, R. I., Lo Leggio, L., Mangiagalli, M., Michalska, K., Moult, J., Najmudin, S., Nardini, M., Nardone, V., Ndeh, D., Nguyen, T.-H., Pintacuda, G., Postel, S., van Raaij, M. J., Roversi, P., Shimon, A., Singh, A. K., Sundberg, E. J., Tars, K., Zitzmann, N. & Schwede, T. Target highlights from the first post-PSI CASP experiment (CASP12, May-August 2016). *Proteins* **86,** 27–50 (2018).

185. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **23,** 47–55 (2014).

186. Engh, R. A., Huber, R. & IUCr. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr. Sect. A Found. Crystallogr.* **47,** 392–400 (1991).

187. Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M. & Karplus, M. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30,** 1545–1614 (2009).

188. Yu, J., Vavrusa, M., Andreani, J., Rey, J., Tufféry, P. & Guerois, R. InterEvDock: a docking server to predict the structure of protein–protein interactions using evolutionary information. *Nucleic Acids Res.* **44,** W542–W549 (2016).

189. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci.* **110,** 15674–9 (2013).

190. Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyrpides, N. C. & Baker, D. Protein structure determination using metagenome sequence data. *Science* **355,** 294–298 (2017).

191. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30,** 1072–1080 (2012).

Vita

SHOURYA SONKAR ROY BURMAN

Ph.D. Candidate
Dept. of Chemical & Biomolecular Engineering
Johns Hopkins University
ssrb@jhu.edu

3400 N. Charles Street
Maryland Hall 306
Baltimore, MD 21218
443-825-2118

EDUCATION

Ph.D. in *Chemical & Biomolecular Engineering* — October 2018 (Expected)
Johns Hopkins University — Baltimore, MD

B.Tech. in *Biological Sciences & Bioengineering* — May 2012
Indian Institute of Technology Kanpur — Kanpur, UP, India

RESEARCH EXPERIENCE

Graduate Research Assistant, Johns Hopkins University — Sep '12–Oct '18
Advisor: Dr. Jeffrey J. Gray, Dept. of Chemical & Biomolecular Engineering
Topic: Computational modeling of interactions of flexible proteins

Summer Undergraduate Research Fellow, California Institute of Technology — May '11–Jul '11
Advisor: Dr. Marianne Bronner, Division of Biology
Topic: Transcriptional regulators of the neural crest gene regulatory network

Undergraduate Research Student, Indian Institute of Technology Kanpur — May '10-Apr '12

Advisor: Dr. Amitabha Bandyopadhyay, Dept. of Biological Sciences & Bioengineering

Topic: Genes governing homing of limb tendons and ligaments

Undergraduate Research Student, Indian Institute of Technology Kanpur — May '09-Apr '10

Advisor: Dr. Dhirendra S. Katti, Dept. of Biological Sciences and Bioengineering

Topic: Polymer-clay nanocomposite mats for tissue engineering

PUBLICATIONS

1. Koehler Leman J, ..., **Roy Burman SS**, ..., Gray JJ, ... & Bonneau R (total 88 authors) "State-of-the art molecular modeling tools in the Rosetta software suite." *In Preparation* (Review article)
2. **Roy Burman SS**, Jeliazkov JR, Labonte JW, Nance ML, Lubin JH, Biswas N & Gray JJ "Predicting protein homomer, heteromer and oligosaccharide interactions using Rosetta in CAPRI rounds 37–45." *In Preparation* (CAPRI special edition issue)
3. **Roy Burman SS**, Yovanno RA & Gray JJ "Flexible backbone assembly and refinement of symmetrical homomeric complexes." *sub judice* (Pre-print copy: https://doi.org/10.1101/409730)
4. Marze NA*, **Roy Burman SS***, Sheffler W & Gray JJ "Efficient flexible backbone protein-protein docking for challenging targets" *Bioinformatics*. Accepted
5. Kalin JH*, Wu M*, Gomez AV*, Song Y*, ..., **Roy Burman SS**, ..., Gray JJ, ..., Schwabe JWR, Mattevi A, Alani RM & Cole PA (total 34 authors) (2018) "Targeting the CoREST complex with dual histone deacetylase and demethylase inhibitors" *Nature Communications*. 9(1), 53
6. Marze NA*, Jeliazkov JR*, **Roy Burman SS**, Boyken SE, DiMaio F & Gray JJ (2017) "Modeling oblong proteins and water-mediated interfaces with RosettaDock in CAPRI rounds 28–35" *Proteins*. 85(3), 479-486
7. Lensink MF, Velankar S, ..., **Roy Burman SS**, Gray JJ, ..., Wodak SJ (total 102 authors) (2016) "Prediction of homo- and hetero-protein complexes by ab-initio and template-based docking: a CASP-CAPRI experiment" *Proteins*. 84(Suppl 1), 323-48

* These authors contributed equally

ORAL PRESENTATIONS

1. "Docking symmetric homomers with flexible-backbone refinement" *RosettaCON*, Leavenworth, WA, August 2018
2. "Flexible-backbone protein docking using motif scoring and large conformational ensembles." *American Institute of Chemical Engineers Annual Meeting*, Minneapolis, MN, November 2017
3. "Flexible-backbone protein docking." *Lectures in Computational Biophysics* at Johns Hopkins University, Baltimore, MD, October 2017
4. "Flexible-backbone protein docking using motif scoring and efficient conformer sampling." *RosettaCON*, Leavenworth, WA, August 2017
5. "Efficient flexible-backbone protein docking." *Regional Computational Biophysics Symposium*, Baltimore, MD, June 2017

SELECTED POSTER PRESENTATIONS

1. "Efficient flexible backbone protein-protein docking for challenging targets." *Biophysical Society Meeting*, San Francisco, CA, February 2018
2. "Efficient flexible protein-protein docking using a diverse ensemble of monomers." *RosettaCON*, Leavenworth, WA, August 2016
3. "Characterization of peptides designed to control crystal nucleation and growth." *Biophysical Society Meeting*, Baltimore, MD, February 2015
4. "Characterization of peptides designed to control calcite growth." *Gordon Research Conference on Biomineralization*, New London, NH, August 2014
5. "Identification of genes essential for attachment of tendons." *Summer Undergraduate Research Grant for Excellence Poster Session*, Kanpur, India, July 2010 **(Awarded Best Poster)**

TEACHING EXPERIENCE

Instructor, *Protein Misfolding Diseases: A Molecular Perspective* 2015
at Johns Hopkins University

Fellow, *Preparing Future Faculty Teaching Academy* 2013-2015
at Johns Hopkins University

Teaching Assistant, *Computational Protein Structure Prediction and Design* 2014
at Johns Hopkins University

Teaching Assistant, *Introduction to Chemical & Biological Process Analysis* 2013
at Johns Hopkins University

AWARDS & HONORS

At the Indian Institute of Technology Kanpur:

Certificate of Merit for Academic Excellence	2011, 2009
Mona and Paramjit Singh Scholarship	2010–2011
Baljit and Nirmal Dhindsa Scholarship	2008–10
Nitish Thakor Scholarship	2008–2009

ACTIVITIES & OUTREACH

- Wrote tutorials for Rosetta Molecular Modeling Suite. (2016)
- Led the Johns Hopkins University's effort to digitize a guide to living in Baltimore as Graduate Representative Organization Guide Chair. (2015-2016)
- Mentored students in Margaret Brent Elementary School to engineer toy solutions to local problems as a part of the STEM Achievement in Baltimore Elementary Schools Program. (2015)
- Provided one-on-one tutoring to local adults seeking a high school-equivalent degree as a part of Johns Hopkins GED Prep. (2013-2015)}
- Coordinated new student orientation, mentored students on academic probation, and organized mental health workshops as Assistant Coordinator of the Counselling Service at Indian Institute of Technology, Kanpur. (2010-2011)